uc3m | Universidad Carlos III de Madrid

Grado Universitario en Ingeniería Informática
2019-2020

Trabajo Fin de Grado

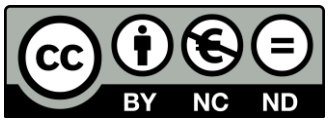
"Data Science explotación y análisis de datos de la calidad del aire en la ciudad de Madrid"

Daniel García Sousa

Tutor

Juan Pedro Llerena Caña

Colmenarejo, 2020



RESUMEN

En el presente estudio se pretende demostrar el papel que puede jugar el análisis de datos y la inteligencia artificial en el control de la contaminación y la meteorología, reduciendo su posible impacto en la salud de las personas y en el medio ambiente.

Más que llegar a una conclusión definitiva, se considera de mayor importancia detallar cuáles son las técnicas y algoritmos existentes y cuáles son las situaciones en las que pueden aplicarse. Desde la clasificación de datos históricos para identificar patrones de comportamiento o tendencias, hasta la predicción de niveles de contaminación, todas estas situaciones son aplicables con los medios de los que se disponen actualmente.

Para cada una de las técnicas aplicadas se llega a una conclusión, valorando su utilidad o precisión, ilustrándose a través de representaciones gráficas.

Palabras clave

Data Science, Ciencia de Datos, Contaminación Atmosférica, Análisis de Datos, Visualización, Regresión, Clustering, Deep Learning

DEDICATORIA

En primer lugar, quiero dar las gracias a Esther, mi compañera en la carrera y en la vida, por todo su apoyo, ayuda y consejo sin los cuales yo no habría llegado a donde estoy ahora mismo. Gracias por absolutamente todo.

También agradecer a mi tutor Juan Pedro y mi tutor "no oficial" Jesús, por los consejos y ayuda y, sobre todo, por toda la paciencia que han mostrado conmigo, que no ha sido poca.

No puedo no mencionar a todos los amigos que he hecho a lo largo de la carrera, desde los dos que me llevo de la primera generación hasta los que llegaron después con los años y con la beca de la universidad. La universidad no habría sido lo mismo sin vosotros.

Por último pero no menos importante, agradecer a mis padres todo su apoyo y confianza en mí, a pesar de los altibajos sufridos sobre todo en los primeros años de universidad.

INDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Motivación del Trabajo

Actualmente, el impacto del ser humano en la naturaleza es una realidad cada vez más grave y que puede darse de muchas formas distintas. La contaminación del aire en especial es una de las mayores preocupaciones por la forma en la que afecta a la salud de las personas, provocando desde asma hasta cáncer en los casos más extremos. La Organización Mundial de la Salud estima que 9 de cada 10 personas en el mundo respira aire contaminado, lo que supone una pérdida de 7 millones de vidas al año [1].

Por ello, un gran número de organismos gubernamentales y universidades han publicado estudios sobre el impacto de estos contaminantes en la salud y su evolución a través del tiempo. Como es el caso del estudio encargado por los organismos de la Unión Europea [2] que analizan los niveles de contaminación de varios países y el número de muertes relacionadas.

Este no es el único problema asociado a la contaminación atmosférica ya que el dióxido de carbono (CO2) es el principal responsable del cambio climático junto con el metano (CH4) y el óxido nitroso (N2O), cuyos niveles se han disparado en la actualidad. En el caso del CO2, el más abundante, en la época de la Revolución Industrial contaba con niveles de 280 partes por millón, llegando en la actualidad a 415 ppm, cifra que puede comprobarse a tiempo real a través de la página oficial de la NASA https://climate.nasa.gov.

Para controlar esta situación y reducir las emisiones de estos contaminantes, en diciembre de 2015 se firmó el Acuerdo de París [3], un tratado global para limitar los efectos del cambio climático e intentar limitar la subida de temperatura global a 1,5 ºC.

Con el presente documento se pretende estudiar cómo ha ido evolucionando la emisión de los distintos contaminantes en la ciudad de Madrid y comprobar si guardan alguna relación entre ellos o con factores meteorológicos aplicando las técnicas y algoritmos más populares de análisis de datos e inteligencia artificial.

1.2. Planteamiento del problema

A continuación, se explica el planteamiento del problema, desde qué punto se parte y cuáles son las técnicas que se aplican.

Partiendo de los problemas asociados a la contaminación atmosférica presentados en el apartado anterior, es posible utilizar técnicas de *Data Science* tanto para ampliar los conocimientos de los que se disponen en la actualidad analizando los datos históricos

distintas magnitudes, como para ser capaz de prever futuros escenarios de contaminación y actuar en consecuencia antes de que se produzcan.

La aplicación de estas técnicas de *Data Science* sólo es posible partiendo de los datos, y en este trabajo se han obtenido datos reales de la ciudad de Madrid de varias fuentes oficiales por lo que es necesarios unificarlos en un formato adecuado para su análisis y, desde ese punto, se realiza el estudio y la aplicación de las técnicas necesarias para la extracción de información.

Este estudio se basa en el uso de las mencionadas técnicas para el análisis de datos históricos meteorológicos y de contaminación atmosférica, para la extracción de información y comportamientos, seguido por el uso de algoritmos de *Machine Learning* para comprobar su alcance y utilidad.

1.3. Objetivos

El objetivo principal de este trabajo es ilustrar cómo el análisis de datos y la inteligencia artificial (y, en última instancia, la tecnología) puede llegar a utilizarse para un fin social, como es un estudio de la calidad del aire, de la misma manera que se está aplicando cada vez más en otros ámbitos, como en la medicina a la hora de procesar radiografías y diagnosticar patologías, o en la identificación de los focos de incendios en imágenes por satélite.

En resumen, los objetivos principales son:

I. Mostrar qué datos meteorológicos y de contaminación están a disposición del público, quién los está ofreciendo y qué información útil se puede sacar de ellos.

II. Utilizar las herramientas más populares actualmente para el análisis y exploración y representación de los datos.

III. Ilustrar, a través de herramientas de visualización, un histórico de los niveles de contaminación y meteorológicos y cuál es la tendencia en la actualidad.

IV. Ganar conocimiento y soltura en el lenguaje de programación Python, el más extendido (junto con R) para el análisis de datos y la inteligencia artificial.

V. Tratar de ampliar los criterios de contaminación de la ciudad de Madrid descubriendo patrones en los datos obtenidos.

VI. Explorar la precisión, el alcance y la utilidad de los algoritmos de Machine Learning más usados, tanto de aprendizaje no supervisado como supervisado.

VII. Utilizar técnicas de regresión para anticiparse a los niveles de los contaminantes que puedan llegar a darse y así aplicar protocolos de contaminación antes de que se produzcan.

1.4. Contenido de la memoria

La presente memoria consta de los siguientes apartados:

1. **Introducción**: motivos de la elección de este tema para su estudio y cuáles son los objetivos que se han perseguido a la hora de elaborar el proyecto.

2. **Estado del arte**: explicación y puesta en contexto de las distintas disciplinas, tecnologías, métodos y componentes que se aplican en este proyecto, tanto a nivel tecnológico como socioeconómico.

3. **Origen y preprocesado de los datos**: descripción de los datos en su forma inicial y su origen, así como las transformaciones aplicadas a ellos para lograr un formato más adecuado para su estudio.

4. **Análisis de datos y extracción de información:** información útil extraída de los datos y a través de qué herramientas se ha llevado a cabo.

5. **Resultados obtenidos**: utilizando las técnicas y algoritmos previamente explicados, interpretar los resultados de cada uno de ellos.

6. **Conclusiones y trabajos futuros**: a partir de los resultados obtenidos, contemplar posibles mejoras que podrían implementarse de cara al futuro.

2. ESTADO DEL ARTE

A lo largo de este apartado se expone el estado del arte de las herramientas, ámbitos y disciplinas relacionadas con este proyecto para comprender la solución planteada dentro de su contexto actual.

Se comienza con una introducción realizando una breve búsqueda de trabajos que abarcan el mismo problema que el presente proyecto y cuáles son los procedimientos que siguen.

Se definen las disciplinas relacionadas con el campo de este estudio como son la Ciencia de Datos, el Análisis de Datos y en qué se diferencian. En los siguientes apartados se expondrán conceptos más tecnológicos relativos a las técnicas y algoritmos utilizados en el proyecto y cómo se han llevado a cabo sus correspondientes implementaciones.

Se termina por el marco legal definido para el uso de los datos utilizados y un contexto socioeconómico del problema de la contaminación atmosférica y su concienciación en distintos ámbitos.

2.1. Estudios previos

Se han realizado numerosos estudios sobre el impacto de la contaminación atmosférica en la salud de la población, en especial lo perjudicial que es el ozono (a nivel del suelo, no en la atmósfera) [4] y las partículas suspendidas en al aire (PM), y pueden verse los informes en varias páginas de administraciones públicas y gobiernos. [5][6]. También es común encontrar informes sobre cómo el calor, las rachas de viento o la lluvia afectan a la calidad del aire, pero de una forma muy genérica sin especificar los contaminantes implicados [7].

Sin embargo, *European Respiratory Society* (ERS) publicó a finales de 2012 un artículo [8] donde analiza, entre otras cosas, el impacto de las condiciones climatológicas en la contaminación del aire, pero sin ilustrar de forma visual las relaciones entre las medidas ni la evolución de estos factores.

Tras esta búsqueda previa, se llega a la conclusión de que los informes sobre el tema del presente documento se agrupan, generalmente, en dos tipos:

- Tipo informativo al público general: en ellos se expone un resumen de las implicaciones de las condiciones atmosféricas en los niveles de contaminación, pero sin entrar al detalle de en qué medida o cuánto han variado en el tiempo.

- Tipo informes de administraciones públicas: en este tipo de informes se analiza minuciosamente los niveles de las magnitudes afectadas y se elaboran conclusiones muy precisas. El problema de este tipo de documentos es que no se

apoya en ayudas visuales, utiliza un lenguaje que no es accesible para el público general.

2.2. Ciencia de los datos

La ciencia de datos (o *Data Science*) es una disciplina que se encarga de extraer conocimiento a través de los datos utilizando conocimientos estadísticos, conocimientos de Machine Learning y comprendiendo las necesidades de negocio o cierto contexto comercial para poder entender cuál es la información más valiosa que podría extraerse.

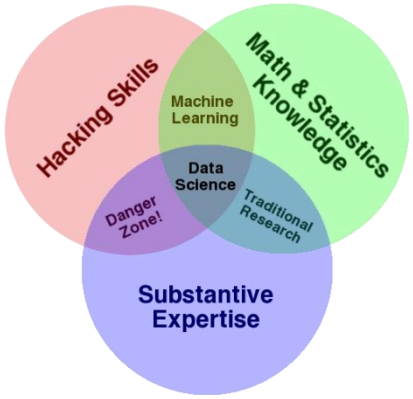


Figura 1. Diagrama de Venn de Data Science [9]

Es un término muy popular hoy en día por la creciente popularidad las técnicas de Big Data en el mundo empresarial y la necesidad de extraer información de interés de volúmenes masivos de datos.

La pregunta que cabría hacerse entonces sería ¿qué es Big Data y por qué es tan popular actualmente?

Se entiende por Big Data un volumen de datos que, por su tamaño, complejidad o dispersión, no puede ser analizado por métodos convencionales. Es aquí donde entran en juego los métodos de análisis de datos para lograr reducir costes, identificar oportunidades de negocio, etc.

La diferencia fundamental entre el análisis de datos y la ciencia de datos es que el análisis trata de responder a una serie de cuestiones específicas sobre un conjunto de datos mientras que la ciencia de datos consiste en averiguar cuáles deberían ser esas cuestiones.

En pocas palabras, la ciencia de datos busca patrones dentro de un volumen de datos que pueden ser estructurados o no estructurados, pueden estar en una base de datos o en varias.

Para ello es necesario utilizar algoritmos de machine learning, inteligencia artificial, herramientas de visualización de datos y herramientas de búsqueda y consulta de datos.

2.3. Machine Learning

Machine Learning es el área dentro de la Inteligencia Artificial que se encarga del desarrollo de programas y algoritmos para que los ordenadores aprendan de manera autónoma en función de unos datos de entrada.

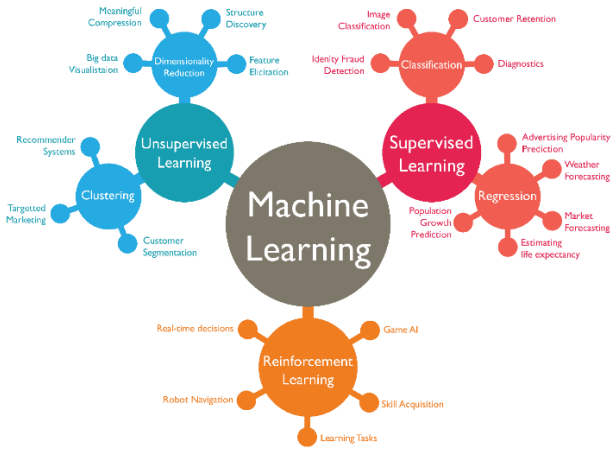


Figura 2. Tipos de Machine Learning [10]

Dentro del Machine Learning se distinguen tres tipos principales de aprendizaje:

- Aprendizaje no supervisado: en este tipo de aprendizaje los datos no tienen asociada una respuesta correcta por lo que la máquina calcula las conclusiones a través de la semejanza entre los datos y buscando patrones entre ellos.

- Aprendizaje supervisado: los datos usados para entrenar a la máquina tienen asociada una "respuesta correcta" para poder generalizar casos nuevos a través de un gran número de ejemplos.

- Aprendizaje de refuerzo: este aprendizaje suele aparecer junto a los dos anteriores tipos de aprendizaje y se diferencia en el que el objetivo es hallar el mejor "camino" a una solución, donde cada acción está penalizada con un coste hasta lograr el objetivo. Este método de aprendizaje en ocasiones es utilizado por los vehículos autónomos para poder averiguar el camino más eficiente.

2.4.1. Aprendizaje no supervisado

Las aplicaciones del aprendizaje no supervisado, en su mayor parte, consisten en el estudio de datos sin estructurar para lograr extraer información, semejanza o patrones de ellos, aunque sus datos son más difusos que los del aprendizaje supervisado.

¿Entonces por qué no aplicar siempre el aprendizaje supervisado si suele da mejores resultados? La razón más simple es que el aprendizaje supervisado requiere que los datos estén etiquetados y estructurados y eso conlleva una gran cantidad de trabajo.

Por otro lado, la capacidad de extraer información de datos no estructurados cobra una gran importancia en la actualidad por la gran cantidad de estos que se generan.

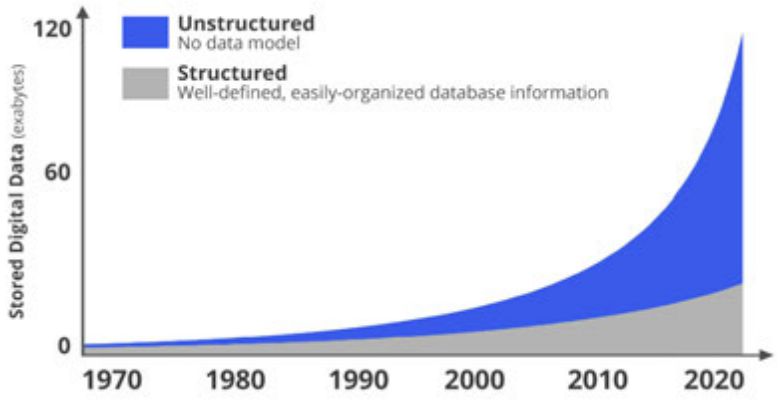


Figura 3. Cantidad de datos estructurados y no estructurados generados [11].

Las técnicas más utilizadas de aprendizaje no supervisado son el agrupamiento o clustering y la reducción de dimensionalidad de datos.

2.4.1.1. Clustering

La clusterización o clustering es una técnica de Machine Learning de aprendizaje no supervisado que consiste en dividir los datos en un cierto número de grupos o clústeres basándose en su similitud.

La similitud entre los datos representados en un espacio de n dimensiones, basándose en el número de variables, se calcula midiendo las distancias entre cada uno de los puntos. En los algoritmos contemplados en este proyecto se ha utilizado la distancia Euclídea para medir la similitud entre los puntos.

- La distancia Euclídea se calcula teniendo en cuenta la localización de los puntos en el espacio. Una de las formas más comunes de obtener esta distancia es calcular la raíz cuadrada de la suma de los cuadrados de las diferencias de sus coordenadas en cada dimensión.

Esta técnica es utilizada para obtener información de un conjunto de datos observando cómo se clasifican dentro de los grupos. Es el propio operador o usuario el que debe validar si la agrupación que se ha llevado a cabo es satisfactoria, es decir, se basa en un criterio subjetivo.

2.4.1.1.1. K-Means

Este algoritmo es probablemente el más conocido de los algoritmos de clustering ya que es muy fácil de ilustrar gráficamente y de implementar.

Los pasos que seguir para implementar este algoritmo son:

- Especificar el número de grupos en los que se quieren dividir los datos.
- Los centros de los grupos o centroides se eligen aleatoriamente.
- Por cada dato, se calcula su distancia a los centroides, se añade al grupo del centroide más próximo y se recalcula el centroide de ese grupo.
- Se repite el paso anterior hasta que los centroides no varíen.

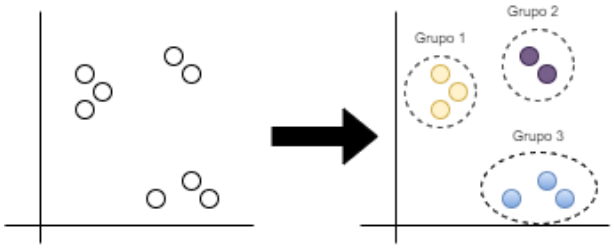


Figura 4. Clustering K-Means

Las ventajas de este algoritmo son que es muy rápido y sencillo, pero es sensible a los valores atípicos u "*outliers*" y los resultados pueden no ser reproducibles. Esto se debe a que los centroides iniciales se inicializan aleatoriamente así que distintas ejecuciones pueden dar distintos resultados.

2.4.1.1.2. Clúster Jerárquico

Otro algoritmo muy popular de clustering es el llamado "clúster jerárquico". Este algoritmo se diferencia del K-Medias en que cada dato se ubica en un clúster distinto y, en cada paso de la ejecución, se identifican los dos clústeres que más cerca están y se juntan en un solo clúster.

En resumen, los pasos del algoritmo son:

- Dado una serie de datos separados cada uno en un clúster distinto.
- Identificar los grupos más cercanos
- Juntarlos en un mismo grupo
- Repetir los pasos anteriores hasta haber agrupado todos los datos en un solo clúster

La manera de representar el resultado de este algoritmo es mediante un dendograma:

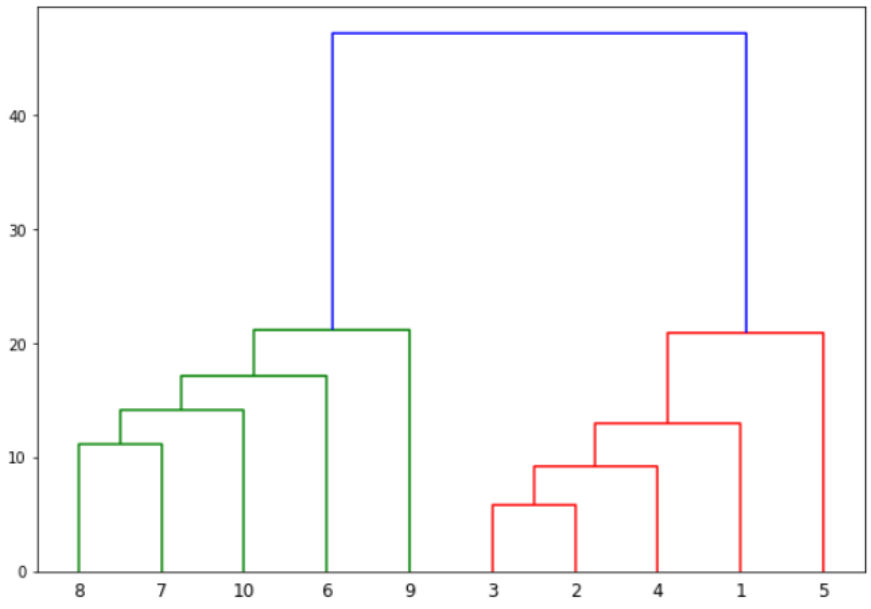


Figura 5. Dendograma [12]

Donde el eje horizontal está compuesto por los distintos datos que se agrupan y el eje vertical muestra la distancia entre estos datos.

Es importante destacar que es necesario especificar una distancia máxima para poder generar un número específico de clústeres. Por ejemplo, en la Figura 5, marcando una distancia máxima de 20 puede decirse que ese conjunto de datos se agrupa en 4 clústeres.

2.4.1.2. Reducción de dimensionalidad

Puede darse el caso en que se disponga de un conjunto de datos con un número de dimensiones mayor de tres y querer representar esas observaciones gráficamente, pero ese alto número de dimensiones lo impide.

Para solucionar este problema existen varias técnicas y algoritmos considerados como aprendizaje no supervisado, siendo el Análisis de Componentes Principales (*Principal Component Analysis* o PCA) y t-SNE los utilizados en este proyecto.

- PCA

La idea principal detrás de PCA es, dado un conjunto de variables o características, escoger las variables que sean más significativas a la hora de calcular el resultado ya que estas características pueden ser linealmente dependientes entre ellas o no ser relevantes para la predicción de la salida.

Para ello los valores de las variables deben estar estandarizados y se generan un número concreto de nuevas características linealmente independientes entre sí que representan un porcentaje de la varianza de los datos iniciales. Estas variables linealmente independientes están formadas por los autovectores de los datos originales y están ordenados en función de la varianza acumulada de los datos que representan. Esto quiere decir que, si los datos se reducen a dos dimensiones, esas dos dimensiones corresponden a los dos autovectores que más datos representan.

Aunque se explica el funcionamiento general del algoritmo de PCA, para más detalle se puede consultar el artículo "*Principal component analysis: a review and recent developments*" [36].

- t-SNE

El algoritmo de machine learning t-SNE (*T-distributed Stochastic Neighbor Embedding*) ha sido diseñado expresamente para la visualización de datos de un alto número de dimensiones en un espacio dimensional reducido y es ampliamente utilizado en varios campos de investigación.

El primer paso del algoritmo es crear una distribución probabilística con pares de datos originales basándose en su similitud, de esta forma los datos similares tienen una mayor probabilidad asignada que los datos dispares. A continuación, traslada esa distribución al espacio dimensional reducido y, de esta forma, se mantiene la información y la robustez de los datos. Una descripción más detallada de la implementación de este algoritmo puede ser consultada en su artículo original [32].

2.4.1. Aprendizaje supervisado

El aprendizaje supervisado, al contrario que el no supervisado, se aplica sobre datos estructurados y en la relación entre entradas y salidas. Está más extendido que el aprendizaje no supervisado por tener un objetivo claro (los resultados de los datos de entrenamiento están ya marcados) y, por tanto, se disponen de métricas para calcular su precisión y rendimiento.

Este tipo de aprendizaje se utiliza en:

- Regresión: la regresión consiste en ajustar los datos de salida basándose en unos datos de entrada. Normalmente esto se aplica en la predicción de valores futuros dados o en el estudio de relaciones entre variables. Este tipo de aprendizaje es el que aplica en el presente documento.

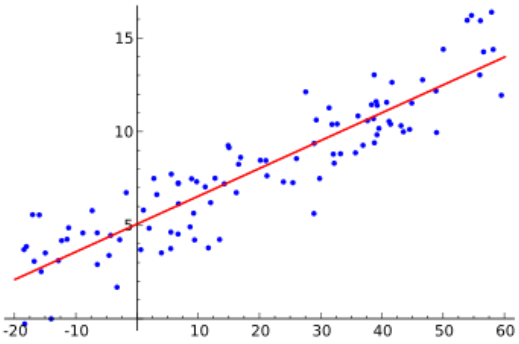


Figura 6. Ejemplo de regresión lineal [13]

- Clasificación: en los problemas de clasificación el objetivo es asignar una categoría basándose en una serie de características o datos de entrada. La diferencia con los problemas de regresión es que el resultado no es un valor numérico sino un vector que muestra a qué categoría pertenece.

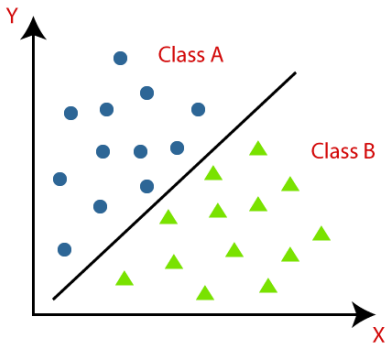


Figura 7. Ejemplo de un problema de clasificación [14]

2.4.1.1. Redes Neuronales Artificiales

Las redes neuronales artificiales son un modelo computacional basado en el funcionamiento de las neuronas biológicas. Este modelo está compuesto por una serie de nodos (las neuronas) conectados entre sí, empezando por la entrada hasta generar una salida.

El ejemplo más sencillo (e ilustrativo) de una red neuronal es el llamado **perceptrón simple** creado por Frank Rosenblatt, una única neurona artificial que genera un valor de salida "**y**" como resultado de una función de activación **f(x)** con un número "**n**" de datos de entrada cada uno con ponderación o peso "**w**".

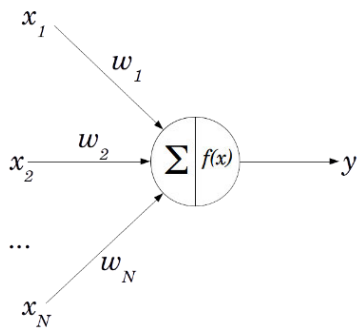


Figura 8. Perceptrón simple [15]

Al estar formado por entradas y salidas binarias el Perceptrón presenta un gran número de limitaciones por lo que han surgido diseños más complejos y que ofrecen mejores resultados, como el **perceptrón multicapa**.

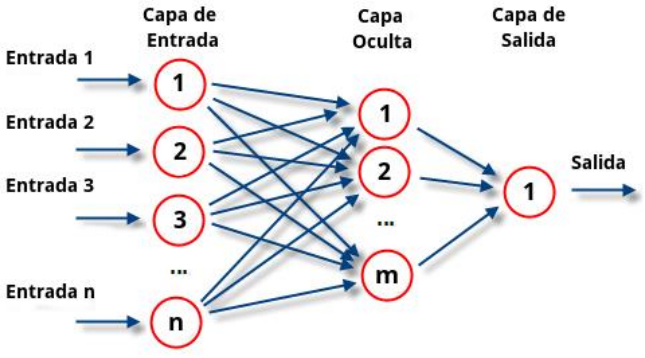


Figura 9. Red Neuronal [16]

Éste consta de un número (n x m) de perceptrones divididos en múltiples capas:

- Capa de entrada: capa de neuronas que reciben los datos de entrada.

- Capa de salida: neurona cuya salida compone el resultado de la red

- Capa oculta: una o más capas de neuronas que reciben las salidas de las capas anteriores y generan las entradas de las siguientes capas. Cuando se habla de **aprendizaje profundo o deep learning** es porque esta capa oculta consta de un gran número de capas, por ejemplo 1000.

¿Por qué son tan famosas las redes neuronales para resolver problemas? Este tipo de Machine Learning obtiene muy buenos resultados con el aprendizaje supervisado por el simple motivo de que, una vez definida la arquitectura de la red neuronal, lo único que hay que hacer es "alimentarlas" en el entrenamiento, es decir, introducir los datos de entrada y especificarle la salida deseada. Siguiendo este proceso, los pesos o ponderaciones entre neuronas se van ajustando para que los cálculos den como resultado la salida esperada.

Para ello las arquitecturas neuronales necesitan una función de coste parametrizada por los parámetros internos de las redes que se pretende minimizar. En el caso de los problemas de regresión es habitual utilizar el error cuadrático medio (*MSE*) mientras que en problemas de clasificación actualmente es habitual utilizar la entropía de Shannon.

Para lograr unos mejores resultados se utiliza el método de entrenamiento de retropropagación o **backpropagation**. Con este método, una vez se introducen los datos de entrada y se obtiene una salida, se compara con la salida deseada y se obtiene un margen de error. Utilizando este margen de error, se va aplicando a las neuronas desde la capa de salida para modificar los pesos capa por capa hasta llegar a la capa de entrada. Utilizando este método la red neuronal se adapta al patrón de los datos del problema en cuestión para poder generar una salida correcta si se le presenta un caso nuevo o inesperado.

Uno de los problemas de añadir demasiadas capas a una red neuronal es el llamado **problema de desvanecimiento de gradiente**. Este problema se produce cuando, al generar la salida y al modificar los pesos de las conexiones de las neuronas en la retropropagación, solo se ven afectadas las primeras capas y las capas ocultas apenas cambian sus pesos. Para solventar este problema existen varios métodos, algunos popularmente conocidos son el ajuste fino de hiperparámetros de entrenamiento [37] o la utilización de otras funciones de activación alternativas como ReLu (*Rectified Linear Unit*) entre otros.

A las redes neuronales es necesario especificar una serie de parámetros, llamados hiperparámetros porque son la configuración definida antes de entrenar la red neuronal y son:

- *Batch size* (tamaño de lote): número de ejemplos con los que se entrena la red neuronal en cada iteración. Un mayor número acorta el tiempo de entrenamiento, pero reduce la precisión del resultado.

- *Epoch* o repeticiones: número de veces que se va a entrenar la red neuronal con los datos de entrenamiento.

- Datos de validación: datos con los que se valida la función de pérdida en cada epoch para comprobar el rendimiento de la red neuronal.

- Optimizador: el optimizador es el método por el cual los pesos de la red neuronal se van modificando con cada iteración. Se ha escogido el optimizador *Adam* por su facilidad de configuración ya que ofrece muy buenos resultados con la configuración por defecto.

- Función de pérdida (conocido como *loss*): la función de perdida es la que determina el error entre la predicción de la red neuronal y el valor esperado, por tanto, el objetivo es minimizar dicho valor. Se ha elegido la función "error medio cuadrático" porque es muy utilizada en los problemas en los que se debe predecir un valor cuantitativo [28].

2.4.1.2. LSTM

Para entender qué son las redes neuronales LSTM (Long Short-Term Memory) primero es necesario entender qué es una **Red Neuronal Recurrente (RNN)**.

Una Red Neuronal Recurrente es aquella que, aparte de generar una salida hacia la siguiente capa, tiene retroalimentación. Esto se traduce en que su salida en un determinado instante de tiempo T depende de su propia salida en el momento T – 1.

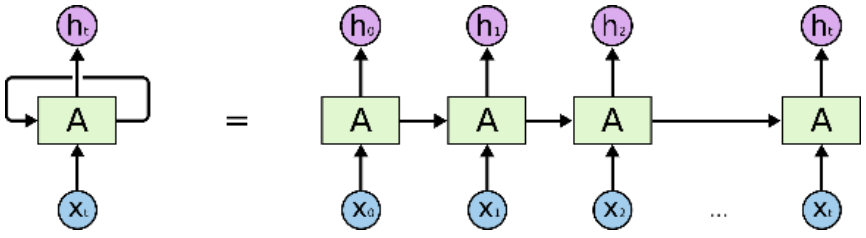


Figura 10. RNN "desplegada" [17]

Este tipo de red neuronal es muy utilizada en estudios de series temporales o de procesamiento de lenguaje, en todo problema que consiste en una secuencia de eventos.

El problema de estas redes es que, si es un periodo de tiempo muy largo, sufren el problema del desvanecimiento de gradiente. ¿Por qué les pasa esto? Explicado en términos coloquiales: estas neuronas, al depender de su propia salida, tienen "memoria" y, si el periodo de tiempo es muy grande, se "olvidan" de lo que ha pasado hace mucho tiempo. Para resolver este problema surgen las LSTMs.

La gran diferencia entre una RNN y LSTM radica en su arquitectura interna. Mientras que las RNNs están compuestas por una sola capa, las LSTMs contienen 4 capas que interactúan entre sí.

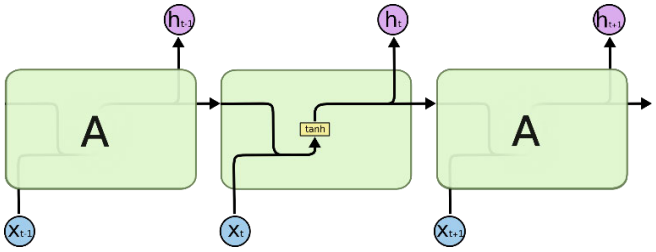


Figura 11. Arquitectura de una RNN [18]

En la Figura 11 se puede ver que una RNN está formada únicamente por una capa con la función de activación de tangente hiperbólica que genera la salida.

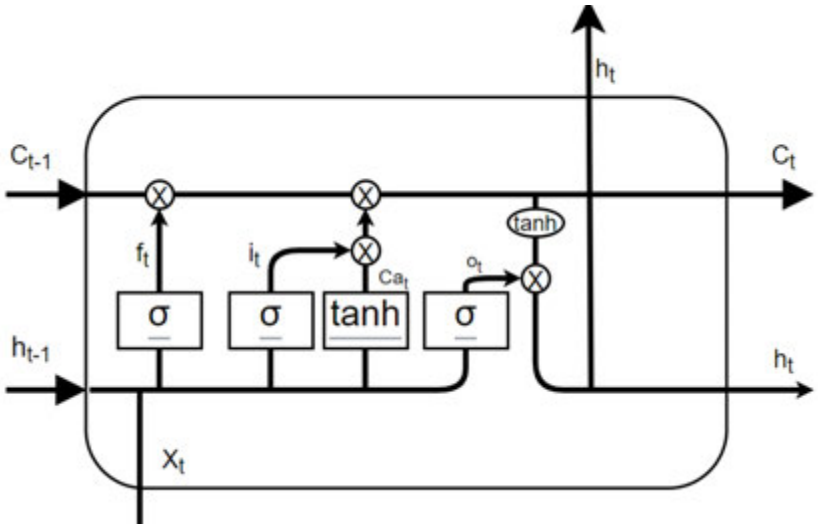


Figura 12. Arquitectura de una LSTM

Por otro lado, en la LSTM se puede ver que se compone de varias capas y operadores:

- **σ**: función sigmoide que genera un valor entre 0 y 1.
- **tanh**: función de tangente hiperbólica.
- **h(t)**: resultado de la neurona en un instante de tiempo "t".

- **C**: es el estado de la neurona y es una de las claves de las LSTMs. Es en este estado donde se mantiene (y borrándose si es necesario) la información necesaria a lo largo de todo el proceso.

- **f(t)**: es la capa de olvidar la información (llamada *Forget Gate Layer*). Utilizando la función sigmoide si genera un 0 no se mantiene la información en ese instante de tiempo "t" y un 1 que se mantenga la totalidad de la información. En términos coloquiales, es en esta capa donde se decide qué información se debe olvidar.

- **i(t) x Ca(t)**: el resultado de la multiplicación de la función sigmoide de i(t) y de la tangente hiperbólica de Ca(t) es la información que es susceptible de añadirse al estado de la neurona C. En términos coloquiales, esta información es la que se deberá tener en cuenta en el futuro.

- **h(t)**: es la salida de la neurona en un instante "t", que también se tendrá en cuenta en el siguiente instante de tiempo cuando vuelva a calcularse todo el proceso.

2.4. Herramientas utilizadas

En este apartado se detallan las herramientas más importantes o que más se han utilizado para la realización de este proyecto. Los criterios a tener en cuenta a la hora de elegir las herramientas han sido:

- Popularidad: las herramientas más populares relacionadas con la ciencia de datos [19] aseguran su eficacia y una comunidad de desarrolladores activa y actualizada.

- Herramientas de código abierto: eligiendo una herramienta de código abierto aumentan las probabilidades de que haya una comunidad de desarrolladores activa y se abarata el coste del proyecto actual.

2.4.1. Python

Python es un lenguaje de programación de alto nivel cuya filosofía es centrarse en la legibilidad del código. Es también un lenguaje multiplataforma, dinámico y multiparadigma, es decir, que soporta programación orientada a objetos, programación imperativa y programación funcional (en menor medida), todo esto por medio de extensiones instalables.

Es el lenguaje de programación más utilizado en análisis y ciencia de datos [20] por su amplia y activa comunidad, la gran cantidad de herramientas estadísticas y matemáticas (Numpy, SciPy, Pandas, etc) y por su facilidad de aprendizaje.

2.4.2. Jupyter Notebooks y Google Colaboratory

Los Jupyter Notebooks son un entorno web de programación interactiva ampliamente utilizados con Python y R en el análisis y ciencia de datos. Esta aplicación fue lanzada en 2015 por la organización sin ánimo de lucro Proyecto Jupyter.

Un Notebook se compone de una serie de celdas de entrada y salida ordenadas que contienen código, texto, contenido multimedia o fórmulas matemáticas. Permite la ejecución de código por bloques independientes, pudiendo volver a ejecutar fragmentos de código sin necesidad de ejecutar todo el que compone el script.

El mayor atractivo para el análisis y ciencia de los datos es la posibilidad de generar reportes, gráficos y resultados en una misma interfaz.

Para la realización de este proyecto se ha utilizado Google Colaboratory, un entorno web basado en los Jupyter Notebooks. Gracias a su integración en Google Drive se puede compartir fácilmente y pone a disposición del usuario una GPU virtual de forma gratuita para las operaciones que tienen un mayor coste computacional, como el entrenamiento de redes neuronales.

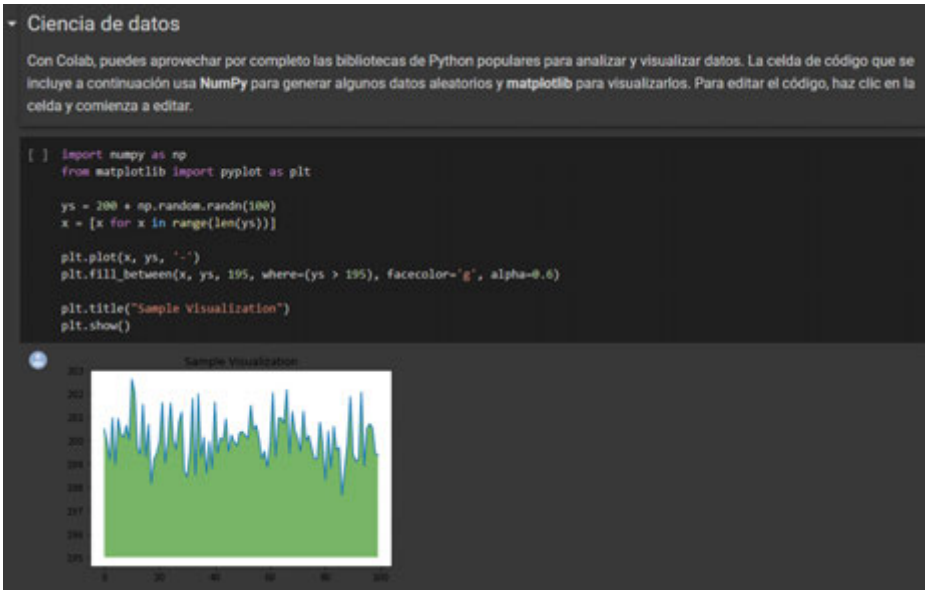


Figura 13. Google Colaboratory [21]

2.5. Marco regulador

Es necesario conocer si existe alguna legislación sobre la utilización de los datos obtenidos para la realización de este estudio. Puesto que los datos provienen del Ayuntamiento de Madrid y de AEMET, se estudiarán las legislaciones de estos dos organismos.

- Política de reutilización de los documentos proporcionados por el Ayuntamiento de Madrid:

"Debe citarse la fuente de los documentos objeto de la reutilización. Esta cita podrá realizarse de la siguiente manera: "Origen de los datos: Ayuntamiento de Madrid (o, en su caso, órgano administrativo, organismo o entidad de que se trate)"

- Política de reutilización de los documentos proporcionados por AEMET:

"Se debe citar a AEMET como fuente de la información objeto de la reutilización en una de las siguientes formas: (...)

En caso, de realizar con ella servicios de valor añadido en base a la información meteorológica y climatológica suministrada por AEMET para su difusión o suministro a terceros, se debe mencionar explícitamente a AEMET como propietaria de dicha información, incluyendo la referencia "Fuente: AEMET" o en su lugar el texto: "Información elaborada utilizando, entre otras, la obtenida de la Agencia Estatal de Meteorología"."

2.6. Contaminación atmosférica

La contaminación atmosférica es la presencia de partículas nocivas en el aire que puede deberse tanto a la acción humana como a un suceso natural y que pueden tener un efecto perjudicial tanto para la salud como para la situación de cambio climático en la que vivimos actualmente.

Es un problema muy grave actualmente porque la industria, el uso de combustibles fósiles y el transporte (entre otras cosas) ha incrementado la cantidad de estas partículas perjudiciales para la salud. Sus efectos se notan ya no solo en los principales focos como las grandes ciudades, donde en los casos más extremos se supera 10 veces el límite [35], sino que están afectando al resto del planeta con efectos como el calentamiento global.

La ciudad de Madrid, siguiendo el Protocolo Anticontaminación [22] se ha dividido en cinco zonas siguiendo varios criterios:

- Distribución de la población.
- Tipología y distribución de las estaciones.
- Estructura de la red viaria.

En el mismo Protocolo se definen cuáles son los valores límites de dióxido de nitrógeno (NO2) y los distintos niveles de actuación:

- **Preaviso**: dos estaciones de una misma zona superan, simultáneamente, 180 µg/m3 durante 2 horas consecutivas o tres estaciones de cualquier zona superan ese nivel durante 3 horas consecutivas.

- **Aviso**: dos estaciones de una misma zona superan, simultáneamente, 200 µg/m3 durante 2 horas consecutivas o tres estaciones de cualquier zona superan ese nivel durante 3 horas consecutivas.

- **Alerta**: tres estaciones de una misma zona (o dos estaciones si se trata de la zona 4) superan, simultáneamente, 400 µg/m3 durante 3 horas consecutivas.

En función de estos niveles de actuación, se definen cinco tipos de escenarios:

- **Escenario 1**: 1 día con superación del nivel de preaviso.

- **Escenario 2**: 2 días consecutivos con superación del nivel de preaviso o 1 día con superación del nivel de aviso.

- **Escenario 3**: 3 días consecutivos con superación del nivel de preaviso o 2 días con superación del nivel de aviso.

- **Escenario 4**: 4 días consecutivos con superación del nivel de aviso.

- **Escenario Alerta**: 1 día de nivel de alerta.

Se utilizarán estos escenarios y niveles de alerta en el desarrollo del proyecto con el objetivo de identificar patrones comunes en los distintos niveles de actuación.

3. ORIGEN Y PREPROCESADO DE LOS DATOS

En el siguiente capítulo se detallan las características iniciales de los datos utilizados y cuáles fueron los pasos seguidos para finalmente unificarlos en un solo formato adecuado para su estudio y análisis.

3.1. Origen de los datos

Los datos que se han sido usados en el presente proyecto provienen de dos fuentes:

- Ayuntamiento de Madrid: fuente de los datos de contaminación en formato de archivo de texto .txt y archivo .csv.

- AEMET: fuente de los datos meteorológicos en formato JSON a través de una API.

3.1.1. Datos horarios de contaminación atmosférica

Los datos horarios de contaminación atmosférica son suministrados por el Ayuntamiento de Madrid en su portal. Estos datos son actualizados periódicamente con las últimas mediciones y están disponibles para todo aquel que quiera acceder a ellos [23].

El formato inicial de los datos varía en función de su fecha:

- Datos anteriores a octubre de 2017

Cada mes de este periodo está recogido en un archivo de texto compuesto por una medición por cada línea con el formato (por ejemplo):

2807900401380217070100005V00004V00004V00004V00004V…

Esta información equivale a la Tabla 1:

PROVINCIA	MUNICIPIO	ESTACIÓN	MAGNITUD	TÉCNICA	PERIODO ANALISIS	AÑO	MES	DIA	DATO	CÓDIGO DE VALIDACIÓN
28	079	004	01	38	02	17	07	01	00005	V

Tabla 1. Representación inicial datos horarios antiguos de contaminación

- Datos posteriores a octubre de 2017

Cada mes de este periodo está recogido en un archivo CSV (*Comma Separated Value*) compuesto por una medición por cada línea con el formato:

28,079,004,01,38,02,2019,01,01,00023, V,00045, V,00028, V,00037, V, …

Esta información equivale a la Tabla 2:

PROVINCIA	MUNICIPIO	ESTACIÓN	MAGNITUD	TÉCNICA	PERIODO ANALISIS	AÑO	MES	DIA	DATO	CÓDIGO DE VALIDACIÓN
28	079	004	01	38	02	2019	01	01	00023	V

Tabla 2. Representación inicial datos horarios actuales de contaminación

3.1.2. Datos meteorológicos diarios

Los datos meteorológicos diarios provienen de la Agencia Estatal de Meteorología (AEMET) y son obtenidos a través del portal de datos *OpenData* [24] utilizando la API REST que ponen a disposición de los interesados.

Con el objeto de este proyecto, se han creado peticiones HTTP para obtener los datos meteorológicos del periodo entre los años 2001 y marzo de 2020. Un ejemplo de los datos obtenidos se muestra en la Tabla 3:

Fecha	T-Media	T-Max	T-Min	Precipitaciones	Sol	Velocidad Viento	Presión Máxima	Presión Mínima
2001-01-02	8,8	11,2	6,3	0,1	4,3	4,4	939,7	931,0

Tabla 3. Muestra de datos meteorológicos

Aunque AEMET ofrece los datos de todas las estaciones meteorológicas de España, para la realización de este proyecto se ha elegido obtener los datos de la estación meteorológica del Retiro en Madrid (código de estación: 3195) ya que se ubica en el centro de la ciudad de Madrid y en este trabajo se considera que representan los datos meteorológicos generales de toda la ciudad.

3.1.3. Estaciones de medición de calidad del aire

Para la detección de alertas siguiendo los criterios definidos por el Ayuntamiento de Madrid, se ha creado un archivo .csv con la información de las estaciones de medición de la calidad del aire indicada en la página web del ayuntamiento de Madrid [25].

Código Área	Nombre Estación	Código Estación	Fecha creación
1	Escuelas Aguirre	8	-
1	Castellana	48	01/06/2010

Tabla 4. Muestra de datos de estaciones de calidad del aire

En especial se hará uso del código de área para poder distinguir la localización de las estaciones, identificadas por su código de estación. Si una estación ha sido dada de baja antes de la fecha de la realización de este proyecto, en la fecha de creación estará indicado con el valor "BAJA".

3.2. Preprocesado de los datos

A continuación, se detallan los pasos seguidos, en cada uno de los conjuntos de datos, para transformar los datos en un formato más adecuado para su estudio y análisis.

3.2.1. Datos horarios de contaminación atmosférica

El primer paso para transformar los datos atmosféricos en crudo en un formato más descriptivo es unificar ambos formatos en un único marco de datos (o *dataframe*) en el que cada medida de tiempo tenga asociado una medición:

Provincia	Municipio	Estación	Magnitud	AA	MM	DD	HH	Dato	Validado
28	79	4	1	2018	4	1	1	2	V

Tabla 5. Formato unificado de los datos de contaminación atmosférica

Partiendo de esta base, se pueden llegar a la conclusión de que las columnas "Provincia" y "Municipio" son prescindibles porque el alcance de este proyecto se centra en la ciudad de Madrid.

Por otro lado, la columna "Validado" nos indica si el dato que representa la fila ha sido validado por la "Dirección General de Sostenibilidad y Control Ambiental Subdirección General de Sostenibilidad" del Ayuntamiento de Madrid.

Una forma de comprobar si estos datos son consistentes es averiguar qué porcentaje de estos han sido validados. Como resulta que los datos no validados sólo representan aproximadamente un 0.97% se puede afirmar que los datos son consistentes y, por ello, se puede prescindir de dicha columna.

Teniendo en cuenta estas consideraciones, sustituyendo el valor de la columna "Magnitud" por el contaminante correspondiente (Anexo II del fichero de descripción de los datos de contaminación de Madrid [25]) y juntando las columnas de año (AA), mes (MM), día (DD) y hora (HH) en una nueva columna "Fecha", el formato del dataframe es el mostrado en la Tabla 6:

Estación	Magnitud	Dato	Fecha
2	SO2	12,73	2001-01-01 1:00:00
2	CO	0,93	2001-01-01 1:00:00

Tabla 6. Formato datos contaminación atmosférica necesarios

Para ilustrar cuales son las relaciones entre las distintas mediciones en un mismo instante tiempo, es necesario ordenar estos datos en función de la fecha que se tomaron y de su estación, pero ¿son necesarias todas las magnitudes?

Consultando los estándares de la Unión Europea de la calidad del aire [26] se indica cuáles son las principales magnitudes que se van a mantener. Si ordenamos los datos resultantes por fecha y estación, se obtiene prácticamente el formato de datos definitivo que se va a utilizar exceptuando los datos que se vayan a obtener a partir de estos. Se utilizará como índice conjunto la fecha y la estación.

Fecha	Estación	SO2	NO2	PM10	O3	PM2.5
2004-02-10 12:00:00	1	24.870001	80.400002	38.450001	17.670000	14.83
	6	18.410000	75.580002	40.799999	16.719999	17.24
	15	0,93	73.739998	50.930000	22.770000	13.40

Tabla 7. Formato datos contaminación atmosférica necesarios

3.2.2. Datos de protocolos anticontaminación

Utilizando los datos de la Tabla 7 y aplicando la información del Ayuntamiento de Madrid, se puede calcular los distintos escenarios y alertas que se aplicaron o se aplicarían (en fechas anteriores a la definición real de estos criterios).

El primer paso es identificar el área a la que pertenece cada estación utilizando el archivo creado con este objetivo y mantener únicamente las estaciones que continúan vigentes hoy en día. También se eliminan todas las magnitudes excepto el dióxido de nitrógeno (NO2), pues es lo que determina los niveles de alerta.

Fecha	Estación	Área	NO2
2001-01-01 00:00:00	4	1	50.240002
	8	1	67.120003
	11	1	49.830002

Tabla 8. Formato de datos de contaminación tras identificar el área

A continuación, tal y como se explica en la sección 2.6 de este documento, se comparan los valores de cada estación con los obtenidos en las horas anteriores y se agrupan esos niveles por área. Teniendo los distintos niveles de contaminación por hora y por área, se pueden calcular qué escenarios han cumplido las condiciones para activarse cada día.

3.2.3. Datos diarios atmosféricos

Los datos atmosféricos son proporcionados por AEMET en su portal de datos abiertos obtenidos por medio de peticiones HTTP a través de su API. Estos datos se reciben en un JSON que, trasladándolos a un dataframe y ordenándolos por fecha, están en su formato definitivo como se muestran en la Tabla 3.

Antes de continuar con la limpieza y ordenamiento de datos, es recomendable comprobar si estos datos están completos o, por el contrario, faltan demasiados datos. Gracias a la librería "pandas" podemos obtener esta medida con una simple línea de código:

Columna	Porcentaje de datos vacíos
FECHA	0
T-MEDIA	0.043234
T-MAX	0.043234
T-MIN	0.043234
PRECIP	0
SOL	66.335207
VEL-VIENTO	5.778931
PRES-MAX	0.807033
PRES-MIN	0.792621

Tabla 9. Porcentaje de datos vacíos en los datos atmosféricos

Aunque el porcentaje de falta de datos admisible en un problema es algo subjetivo, se va a prescindir de la información relativa al Sol por faltar más de un 66% de las medidas. Por otro lado, se va a mantener, al menos por el momento, la información de la presión atmosférica, aunque juntándolas en una columna con la presión atmosférica media. Se eliminan también las columnas de temperaturas máximas y mínimas para evitar redundancia con la temperatura media.

Fecha	T-MEDIA	PRECIP	VEL-VIENTO	PRES-MEDIA
2001-01-01	8.2	5.2	2.2	934.00

Tabla 10. Muestra de datos meteorológicos definitivos

En el resto de las magnitudes, al ser un número muy reducido de datos vacíos, se rellenan los datos que faltan con la media de cada columna.

3.2.4. Marco de datos definitivo

Ahora que los distintos datos están definidos y en un formato prácticamente unificado, es el momento de juntarlos en un solo marco de datos. Cabe destacar que los datos de calidad del aire son horarios, pero los atmosféricos y los protocolos de contaminación son diarios, por lo que se usarán de base los datos horarios.

Para unir los datos de contaminación con los atmosféricos se toma como referencia el día en el que se hizo la medición, por lo que cada día tendrá unos datos de contaminación distintos en cada hora, pero los mismos meteorológicos y de protocolo anticontaminación a lo largo de este.

Fecha	Estacion	T-Media	Precip	Vel-Viento	Pres-Media	Escenario	SO2	NO2	PM10	O3	PM2.5
2001-01-01 01:00:00	4	8.2	5.2	2.2	934	0	20.61	50.24	19.65	5.98	0.0
	8	8.2	5.2	2.2	934	0	26.46	67.12	32.35	7.86	0.0

Tabla 11. Muestra conjunta de datos meteorológicos y de contaminación

Este marco de datos el que usará en las siguientes fases del proyecto por lo que se va a explicar cada uno de sus componentes:

- *Fecha* (Índice): momento de la medición de la calidad del aire. Está compuesta por la fecha y la hora.

- *Estación* (Índice): código identificativo de la estación de medición de la calidad del aire.

- *T-Media*: medida numérica que indica la temperatura media a lo largo del día. La medida está en grados centígrados.

- *Precip*: medida numérica que indica la cantidad de precipitaciones en ese día. Se mide en milímetros de agua (mm).

- *Vel-Viento*: medida numérica que indica la velocidad del viento en ese día. Se mide en metros por segundo (m/s).

- *Pres-Media*: medida numérica que indica la presión atmosférica en ese día. Se mide en hectopascales (hPa).

- *Escenario*: medida cualitativa que indica el Escenario del protocolo anticontaminación del Ayuntamiento de Madrid que se aplicó (o se aplicaría, en caso de que sea una fecha anterior a la creación de dicho protocolo). Sus posibles valores son del 1 al 5.

- *SO2*: medida numérica que indica la cantidad de Dióxido de Azufre que se emitió a la hora y día indicados. Se mide en microgramos por metro cúbico ($\mu g/m^3$).

- *NO2*: medida numérica que indica la cantidad de Dióxido de Nitrógeno que se emitió a la hora y día indicados. Se mide en microgramos por metro cúbico ($\mu g/m^3$).

- *PM10*: medida numérica que indica la cantidad de PM10, partículas sólidas o líquidas presentes en la atmósfera con un diámetro comprendido entre 2.5 y 10 micrómetros, que se emitió a la hora y día indicados. Se mide en microgramos por metro cúbico ($\mu g/m^3$).

- *O3*: medida numérica que indica la cantidad de Ozono que se emitió a la hora y día indicados. Se mide en microgramos por metro cúbico ($\mu g/m^3$).

- *PM2.5*: medida numérica que indica la cantidad de PM10, partículas sólidas o líquidas presentes en la atmósfera con un diámetro inferior a 2.5 micrómetros, que se emitió a la hora y día indicados. Se mide en microgramos por metro cúbico ($\mu g/m^3$).

4. ANÁLISIS DE DATOS Y EXTRACCIÓN DE LA INFORMACIÓN

A continuación, se detalla el análisis de exploración de datos realizado, detallando las características que se utilizan en los siguientes apartados. Para este fin se hace uso de herramientas de análisis de datos, de análisis estadístico y de visualización.

Empezando por una preparación de los datos para su análisis, se visualizan a continuación los datos históricos de cada una de las magnitudes para comprobar si siguen alguna tendencia.

4.1. Preparación de los datos

Antes de visualizar y analizar los datos, es recomendable comparar si hay datos ausentes o valores atípicos y, si existen, corregirlos de una manera adecuada.

4.1.1. Datos ausentes

Como se indica en apartados anteriores, para el análisis se parte del formato de datos de la Tabla 11. Sin embargo, como se ve en la Tabla 9, hay mediciones de las que no disponemos datos y esto puede pasar por varios motivos:

- Errores en los instrumentos de medición.
- Problemas al procesar los datos.
- Error humano en el momento que se validaron esos datos.

En el caso de que falte un gran número de datos, lo más recomendable es eliminar la magnitud entera pero no hace falta llegar a ese punto si son un número reducido. El primer paso, por tanto, es saber cuántos datos están incompletos:

Columna	Porcentaje de datos vacíos
T-MEDIA	0.0
PRECIP	0.0
VEL-VIENTO	0.0
PRES-MEDIA	0.0
ESCENARIO	0.0
SO2	0.0
NO2	0.0
PM10	0.0
O3	0.0
PM2.5	0.0

Tabla 12. Porcentaje de datos vacíos en el marco de datos base del análisis

Parece que no falta ningún dato, lo cual tiene cierto sentido porque los datos meteorológicos ya fueron corregidos con la media de las columnas en la sección 3.2.3 y los datos de contaminación atmosférica se han sacado de un archivo cuyos datos no validados fueron despreciados en la sección 3.2.1.

A pesar de tener la completa seguridad de que a las mediciones no les pueden faltar datos, es una buena práctica comprobarlo para cerciorarse de que es así.

4.1.2. Estandarización de los datos

Con el fin de poder comparar las distintas magnitudes entre sí es necesario estandarizar los datos para que todos estén en el mismo rango. Por ejemplo, si se quisiera comparar la evolución de la temperatura media (con un valor máximo de 40 aproximadamente) y el ozono (con un valor máximo de 16162 aproximadamente), en el gráfico resultante no aportaría ninguna información porque son medidas que no pueden compararse entre sí.

Para solucionar este problema se aplica la estandarización de los datos, con lo que la media de todas las magnitudes se centra en 0 y la desviación estándar se fija en 1. Haciendo uso de la librería de Python *scikit-learn*, esta estandarización se puede aplicar en una sola línea de código. Es importante aclarar que sólo se aplica sobre magnitudes numéricas, es decir que las medidas cualitativas como el Escenario no se verán afectadas.

Cabe destacar que, a partir de este momento, todos las técnicas y algoritmos se aplican sobre los datos de una única estación, la ubicada en Casa de Campo, por ser una estación activa desde que se tiene constancia, ubicada en el centro de la ciudad de Madrid y con una gran calidad de mediciones de contaminantes. Se utilizarán, también, los datos del

año 2003 en adelante, pues es a partir de este momento cuando se estabilizan las mediciones de la mayoría de los contaminantes.

La estandarización es ampliamente utilizada, además de para visualizar distintas variables en una misma escala, para potenciar los resultados de los algoritmos de Machine Learning, tanto el clustering como los algoritmos de Deep Learning.

Fecha	T-Media	Precip	Vel-Viento	Pres-Media	Escenario	SO2	NO2	PM10	O3	PM2.5
2003-01-01 01:00:00	-0.809	-0.068	0.024	0.803	0	-0.032	-0.977	0.03	0.232	-0.296

Tabla 13. Muestra de datos estandarizados

4.1.3. Datos atípicos

Los valores atípicos de un conjunto de datos se pueden definir como las observaciones que difieren en mucho del resto, lo que hace sospechar de su origen o validez.

Para detectar estos valores atípicos en los datos se utiliza una combinación de dos técnicas: los diagramas de caja y la unidad tipificada o *z-score*.

En primer lugar, se utilizan los diagramas de caja para representar los cuartiles de cada una de las magnitudes y los valores atípicos.

- *T-Media*: no hay ningún valor atípico.

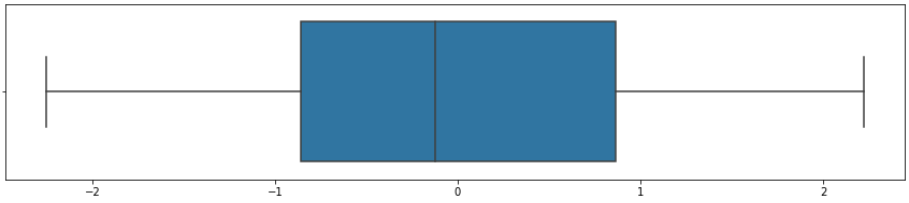


Figura 14. Diagrama de caja de los datos de temperatura media.

- *Vel-viento*: aunque gran parte de los datos están uniformemente distribuidos, existen un número reducido de valores atípicos de valor más alto que los datos acotados.

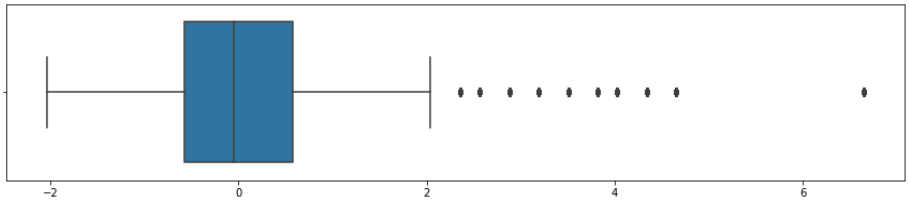


Figura 15. Diagrama de caja de los datos de velocidad del viento.

- *Pres-Media*: mientras que gran parte de los valores se agrupan en torno al valor 0, existen un gran número de valores atípicos de valores más alto y más bajos que los datos acotados.

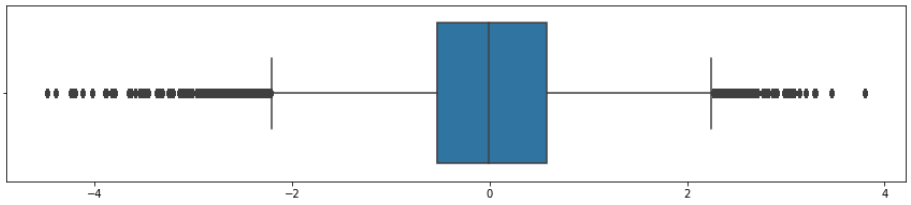


Figura 16. Diagrama de caja de los datos de presión atmosférica media.

- *NO2*:

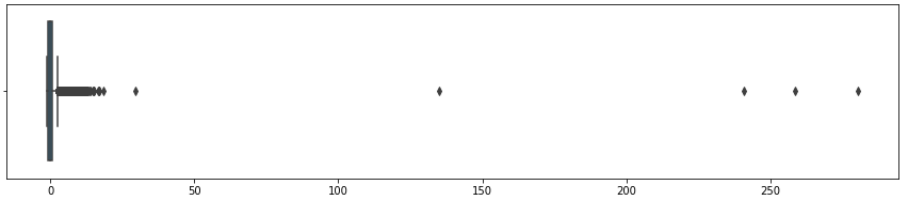


Figura 17. Diagrama de caja de los datos de dióxido de nitrógeno.

Se puede observar que hay muchos datos atípicos en varias magnitudes, tanto meteorológicas como de contaminación, pero no es necesario corregir cada uno de ellos.

Para identificar cuáles son realmente los valores atípicos, se utiliza la unidad tipificada para detectar cuales son los puntos que se alejan de la media de los datos más de +/-3 desviaciones estándar. ¿Por qué justo 3 desviaciones estándar? La unidad tipificada está enfocada, principalmente, para ser aplicada en muestras de datos que sigan una distribución normal o Gaussiana:

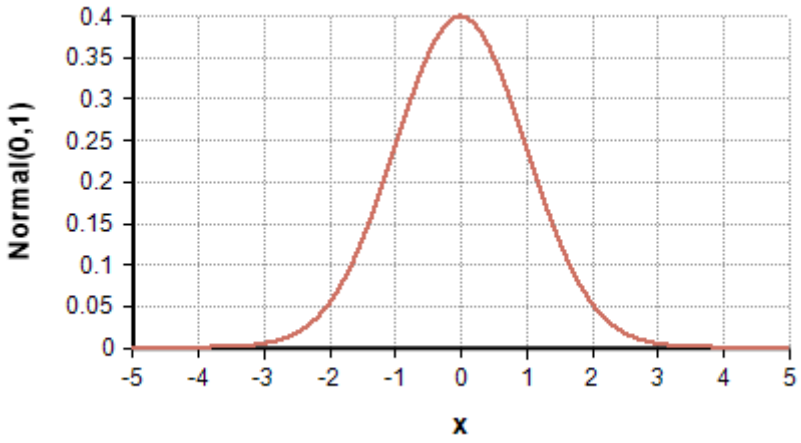


Figura 18. Distribución normal o Gaussiana [27]

Como se puede apreciar en la Figura 18, en una distribución normal gran parte de los datos están centrados en el 0 y luego van disminuyendo a medida que aumenta el valor absoluto de la desviación típica (el eje X) y a partir de 3 desviaciones típicas apenas hay datos, por lo que todo dato que esté más alejado de ese punto, se considera un valor atípico. Pero ese es el caso de una distribución normal, que no de los datos de este trabajo, aunque sí es posible utilizarse como una medida orientativa junto con la representación visual de los diagramas de caja.

Gracias a la librería *SciPy* de Python, la implementación de esta medida es muy sencilla por lo que los pasos a seguir son:

- Detectar los valores con una unidad típica absoluta mayor que 3.
- Sustituir dichos valores por datos ausentes (NaN)
- Sustituir los datos ausentes por el valor máximo de esa magnitud en caso de que el valor atípico fuera mayor que la media o por el valor mínimo en caso de que fuera menor.
- Comprobar gráficamente los diagramas de caja de los datos corregido y, si se considera oportuno, volver a aplicar la unidad tipificada o una corrección manual.

Volviendo a comprobar los diagramas de caja:

- *Vel-viento*: los datos siguen una distribución parecida a antes de eliminar los valores atípicos, aunque de una forma más uniforme.

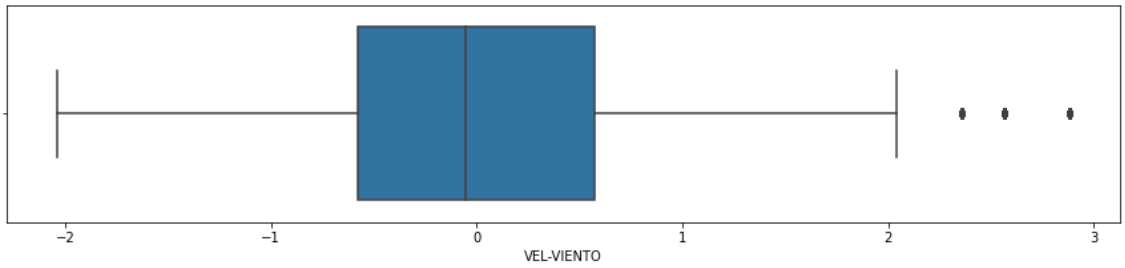


Figura 19. Diagrama de caja de los datos de velocidad del viento corregidos.

- *Pres-Media*: se han reducido el número de valores atípicos tanto mayores como menores que la media.

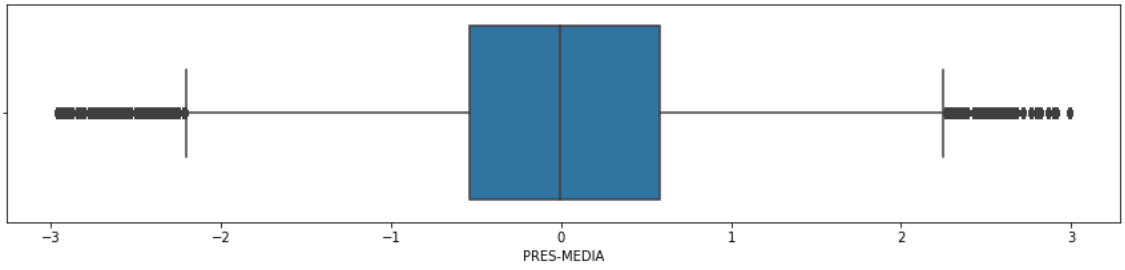


Figura 20. Diagrama de caja de los datos de presión atmosférica media corregidos.

- *NO2*: se han eliminado un gran número de valores atípicos que dificultaban la visualización de los datos.

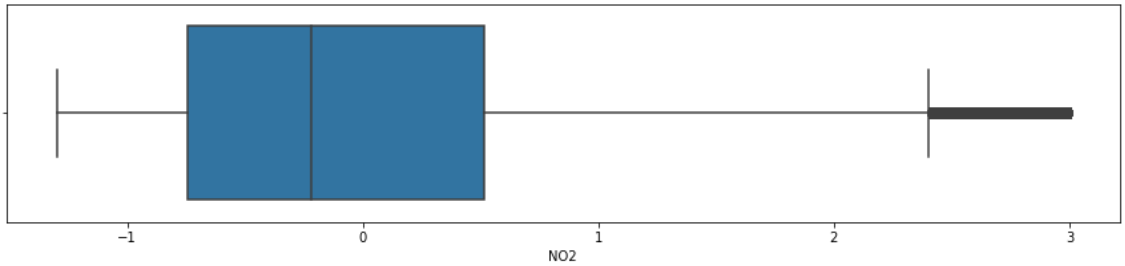


Figura 21. Diagrama de caja de los datos de dióxido de nitrógeno corregidos.

Gracias a la visualización de los diagramas de cajas y la unidad tipificada, se puede comprobar a simple vista que se han reducido los casos más extremos de valores atípicos y se continua con el estudio de las magnitudes a pesar de que algunas presenten ciertos valores extremos.

Cabe destacar que el procedimiento de corrección de valores atípicos se ha realizado sobre todas las magnitudes numéricas si bien solo se han usado las magnitudes mostradas en el presente documento con fines ilustrativos de cómo se ve afectada la dispersión de los datos.

4.2. Visualización de datos históricos

La visualización de los datos es la representación gráfica de la información y puede ser utilizada para facilitar la comprensión, comparación e interpretación de los datos.

Con este objetivo, en los siguientes apartados del presente documento se procede a estudiar el histórico de cada una de las magnitudes para llegar a comprender su comportamiento y evolución a través del tiempo.

Para profundizar en dicho comportamiento, se separan los datos históricos en varios componentes:

- Observed: representación de los datos.
- Trend: tendencia de los datos a través del tiempo.
- Seasonal: los ciclos de repetición a través de la evolución de los datos.
- Residual: variaciones de los datos a través de la tendencia.

4.2.1. Preparación de datos para su visualización

Como pasos previos a la representación gráfica, se crean nuevas columnas enfocadas a facilitar la visualización:

- *Year*: año en el que se tomó la muestra. La utilidad de esta medida es para identificar posibles comportamientos similares cada año, lo que demostraría que los datos son estacionales.

Además, dada la cantidad de datos disponible y la necesidad de plasmarlo en un gráfico, se agruparán los datos en semanas y sustituyendo las magnitudes por la media de dicha semana, de esta forma se reduce la cantidad de datos, pero manteniendo su validez, al menos a la hora de estudiar la tendencia.

4.2.2. Visualización y extracción de información

A continuación, se va a proceder a la representación gráfica de cada una de las magnitudes y al análisis de cada una de ellas.

- *Temperatura Media*

A la hora de estudiar la variación de la temperatura media, la primera hipótesis podría ser que lo valores más altos serán en verano y los más bajos en invierno. Estudiando la tendencia en la Figura 22 a lo largo de los años, se puede afirmar con cierta seguridad que la temperatura sigue un patrón con unos valores mínimos a principios y finales de año (invierno) y unos valores máximos a mitad de año (verano).

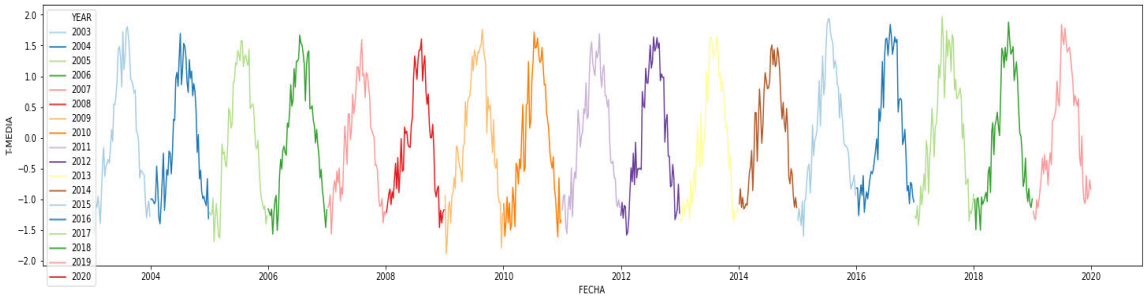


Figura 22. Datos históricos semanales de temperatura media

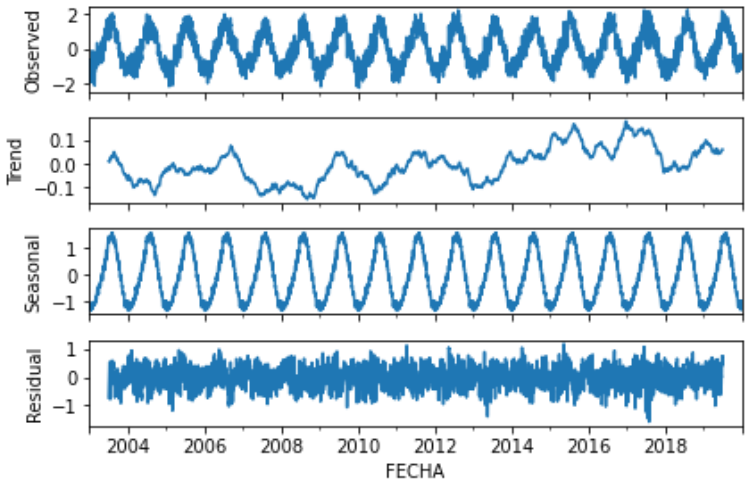


Figura 23. Descomposición estacional de la temperatura media

Observando la tendencia en la Figura 23 se comprueba que la temperatura ha ido subiendo ligeramente desde el año 2014, aproximadamente, hasta la actualidad. Es necesario disponer de un rango de tiempo mucho mayor para poder atribuir este aumento al calentamiento global.

- *Precipitaciones*

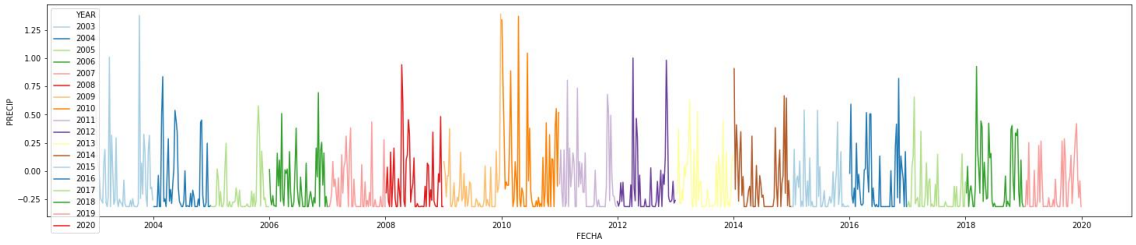


Figura 24. Datos históricos semanales de precipitaciones

Observando el gráfico de la Figura 24 no se puede apreciar a simple vista que haya variado en gran medida la cantidad de precipitaciones, únicamente que los años 2003 y 2010 fueron especialmente lluviosos en Madrid.

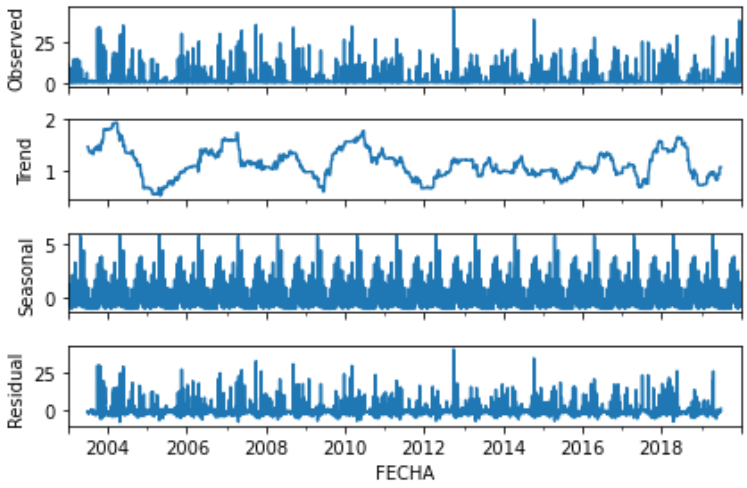


Figura 25. Descomposición estacional de las precipitaciones

- *Velocidad del viento*

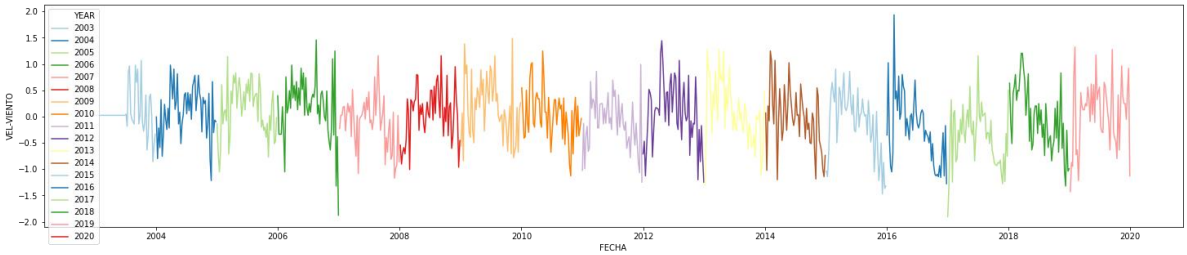


Figura 26. Datos históricos semanales de velocidad de viento

La velocidad media del viento parece mantenerse estable a lo largo de los años, exceptuando el año 2016 con unos niveles anormalmente bajos, con un comportamiento similar año tras año.

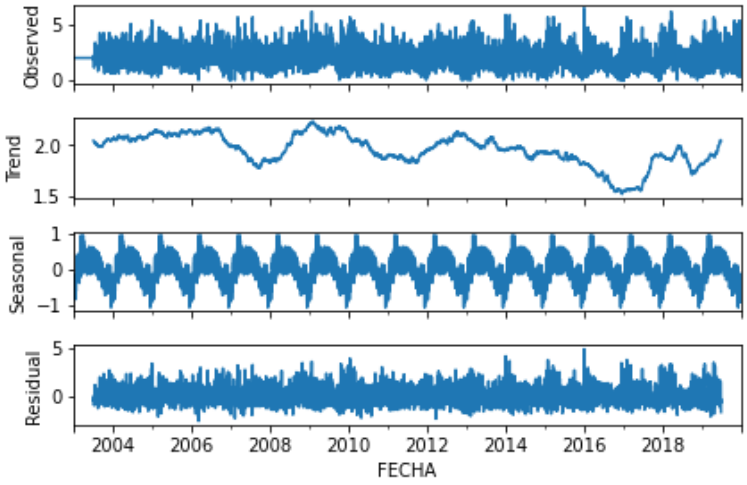


Figura 27. Descomposición estacional de la velocidad del viento

- *Presión atmosférica*

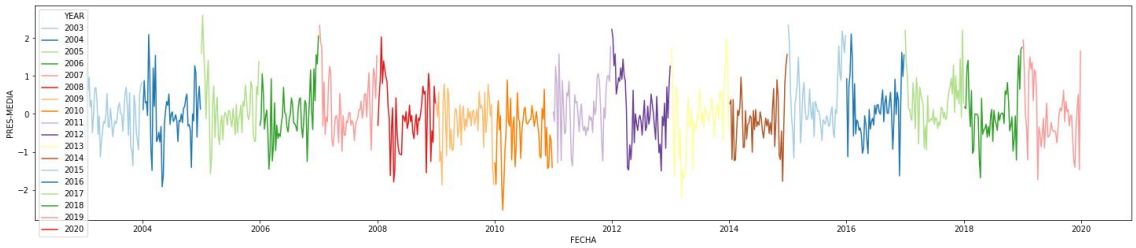


Figura 28. Datos históricos semanales de presión atmosférica

De igual manera que la velocidad del viento, la presión atmosférica parece mantener la misma media con pequeñas variaciones en algunos años.

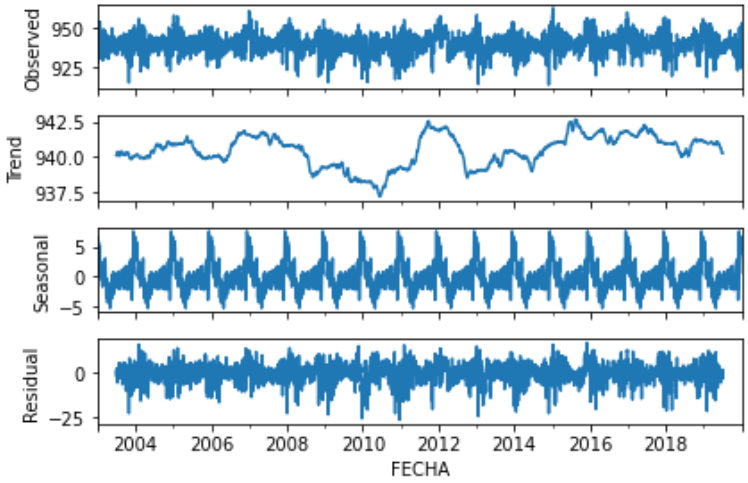


Figura 29. Descomposición estacional de la presión atmosférica media

- *Dióxido de azufre (SO2)*

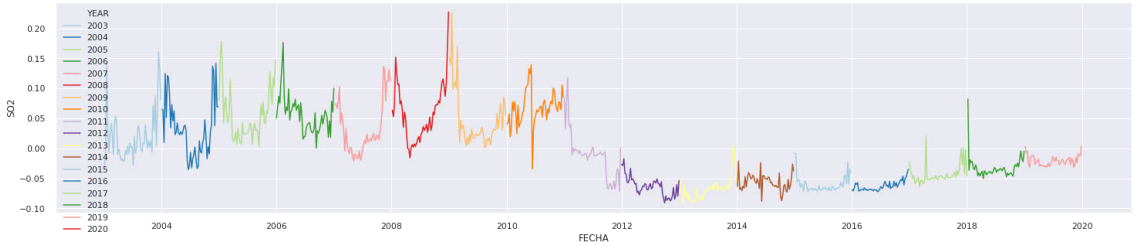


Figura 30. Datos históricos semanales de SO2

Los niveles de SO2 han disminuido considerablemente desde el año 2009 hasta la actualidad y parece tener un componente estacional muy marcado. Estudiando su tendencia de igual manera que con la temperatura media se comprueba su tendencia a reducirse.

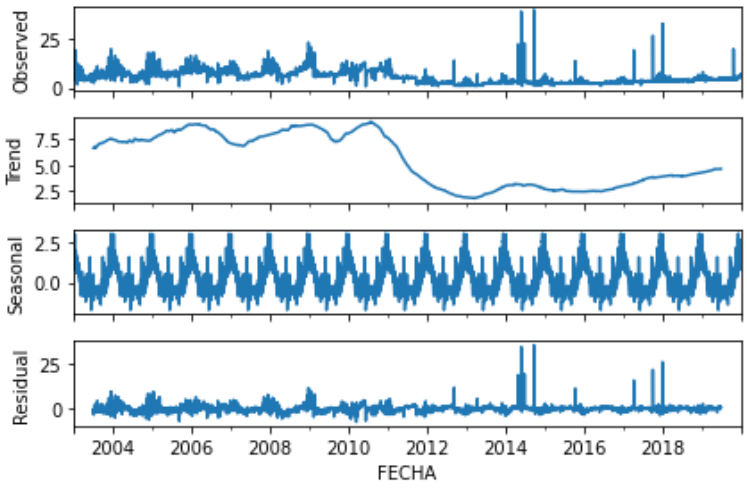


Figura 31. Descomposición estacional del dióxido de azufre

- *Dióxido de nitrógeno (NO2)*

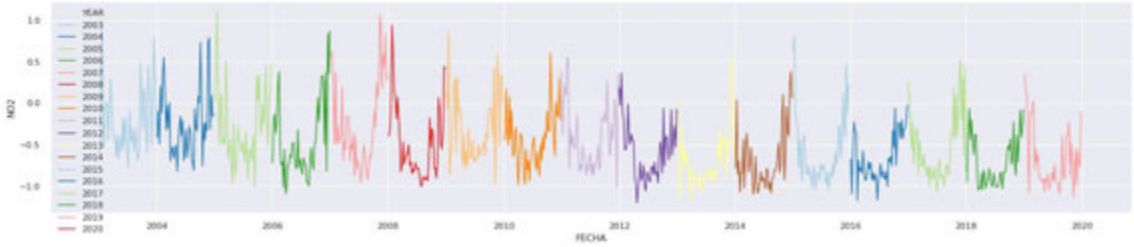


Figura 32. Datos históricos semanales de NO2

Aunque se puede deducir una estacionalidad y una reducción en las emisiones de NO2, no es muy evidente a simple vista. Para comprobar esta estacionalidad, se comprueba el comportamiento de esta magnitud a lo largo de los meses de varios años:

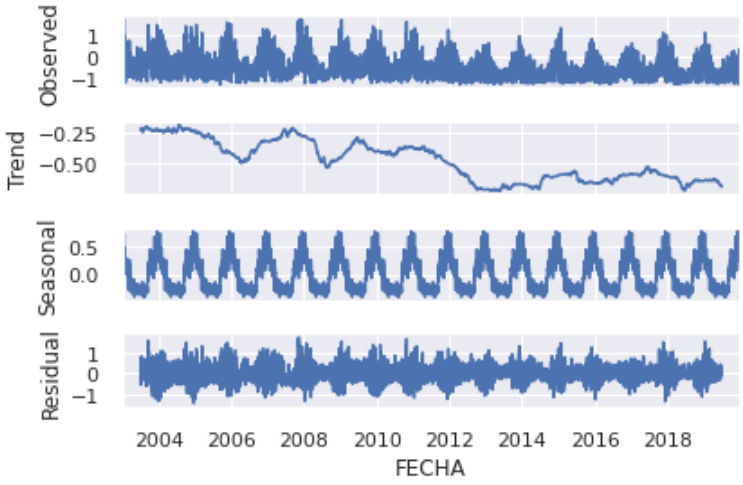


Figura 33. Descomposición estacional del dióxido de nitrógeno

Se puede observar una considerable reducción de emisiones de NO2, probablemente por ser la magnitud más controlada de la ciudad de Madrid al ser la responsable de marcar los criterios de activación de los escenarios anticontaminación.

- *PM10*

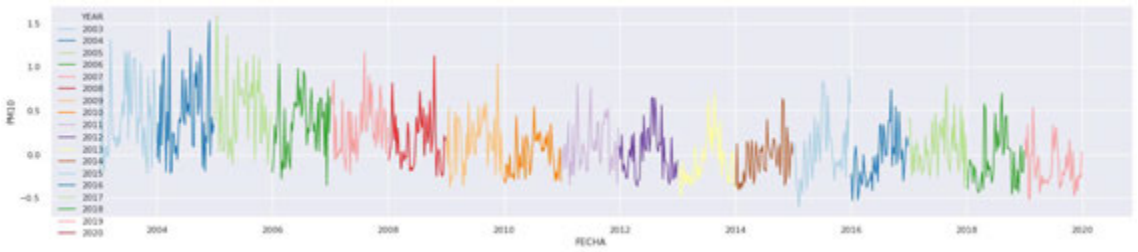


Figura 34. Datos históricos semanales de PM10

No se distingue un patrón claro en el comportamiento de los niveles de PM10, pero se puede analizar la concentración media por cada año y así tener una idea aproximada de las variaciones que ha estado sufriendo.

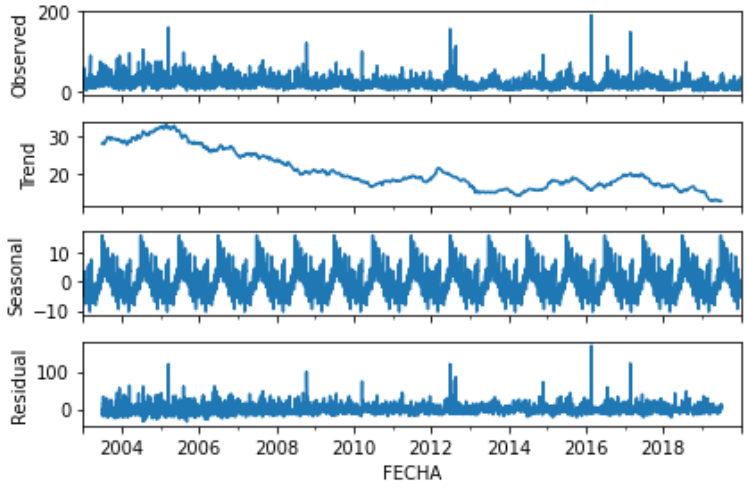


Figura 35. Descomposición estacional de PM10

Cada año registra un nivel menor desde el año 2006 hasta el 2012 donde parece haberse estabilizado hasta la actualidad.

- *O3*

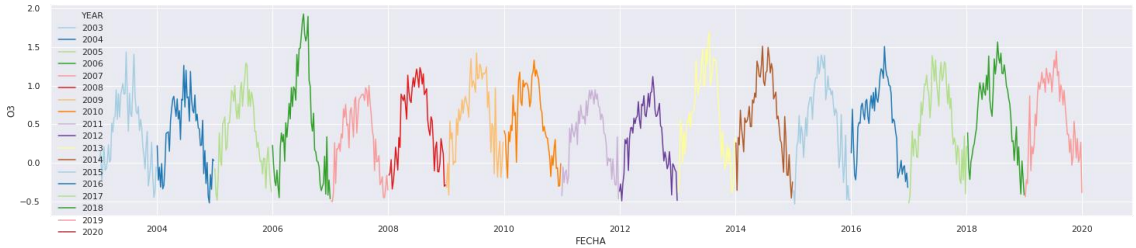


Figura 36. Datos históricos semanales de O3

Otro claro ejemplo de estacionalidad es el ozono y se puede comprobar a simple vista que alcanza sus niveles máximos en verano y que sus niveles anuales han ido aumentando.

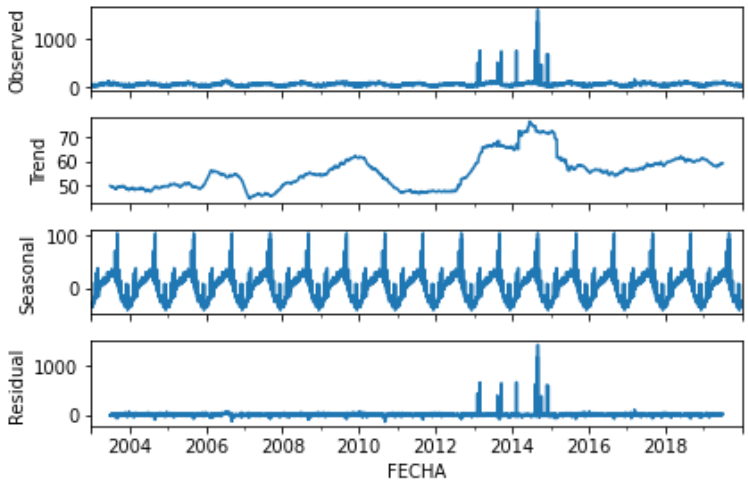


Figura 37. Descomposición estacional del ozono

- *PM2.5*

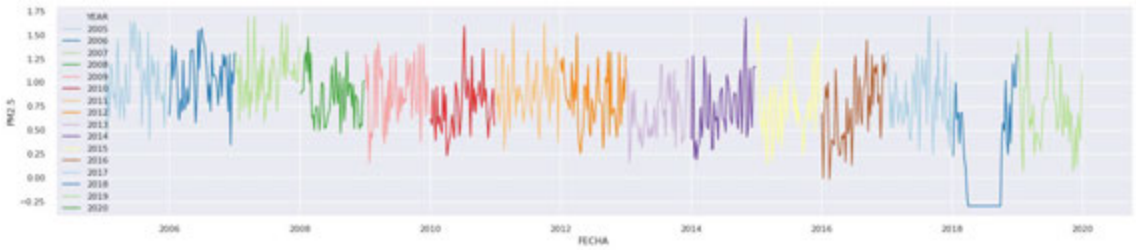


Figura 38. Datos históricos semanales de PM2.5

Al igual que PM10, no se distingue un patrón claro en el comportamiento de los niveles de PM2.5 pero se puede analizar la concentración media por cada año y así tener una idea aproximada de las variaciones que ha estado sufriendo.

Cabe destacar el periodo entre los años 2018-1019 por la evidente bajada de concentración que se puede afirmar con casi total seguridad que se trata de un error de medición.

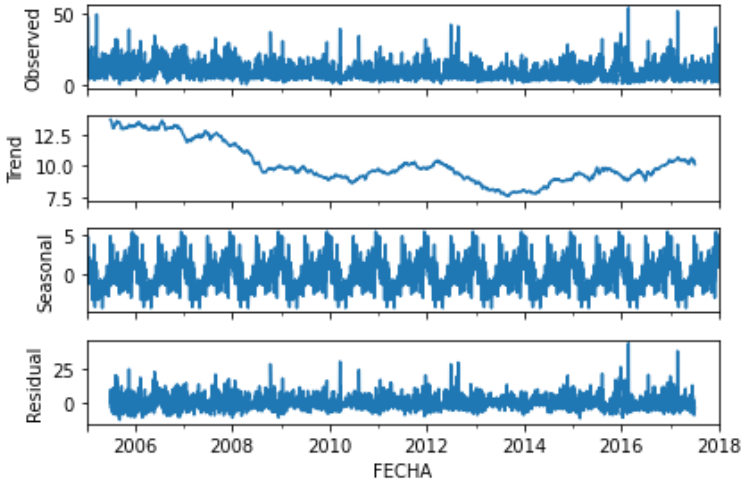


Figura 39. Descomposición estacional de PM2.5

A excepción del bajo nivel del año 2018, sin duda causado por el fallo de medición previamente mencionado, los niveles de concentración de PM2.5 prácticamente no se han reducido desde el año 2006.

4.3. Regresión

Existe un gran número de métodos para predecir los futuros valores de series temporales y en el transcurso de este documento se van a aplicar los más extendidos para comparar sus resultados. Se utilizará un porcentaje de los datos reales para ajustar o entrenar el modelo y los datos restantes para ser comparados con las predicciones.

El error de cada uno de los modelos se mide utilizando el error medio cuadrático (*mean squared error*) entre la predicción y el valor real.

Se van a utilizar dos tipos de predicciones:

- Predicción univariante: los resultados se predicen en función de los valores anteriores de una misma variable. Por ejemplo, los valores de dióxido de nitrógeno dependen de los valores que haya tenido en el pasado.

- Predicción multivariante: los resultados se predicen en función de los valores de n variables. Por ejemplo, los valores de dióxido de nitrógeno se predicen en función de los valores de varias magnitudes, como pueden ser las atmosféricas y los propios valores previos de NO2.

Dado que es el contaminante en el que se basan los escenarios anticontaminación de la ciudad de Madrid, las pruebas se han realizado sobre los datos de dióxido de nitrógeno (NO2).

4.3.1. Correlación lineal

La forma más sencilla de predecir futuros valores de una magnitud es estudiando su relación o correlación lineal con el resto de las variables. Esto indica cuán fuerte aumenta (o disminuye) una magnitud cuando varía otra.

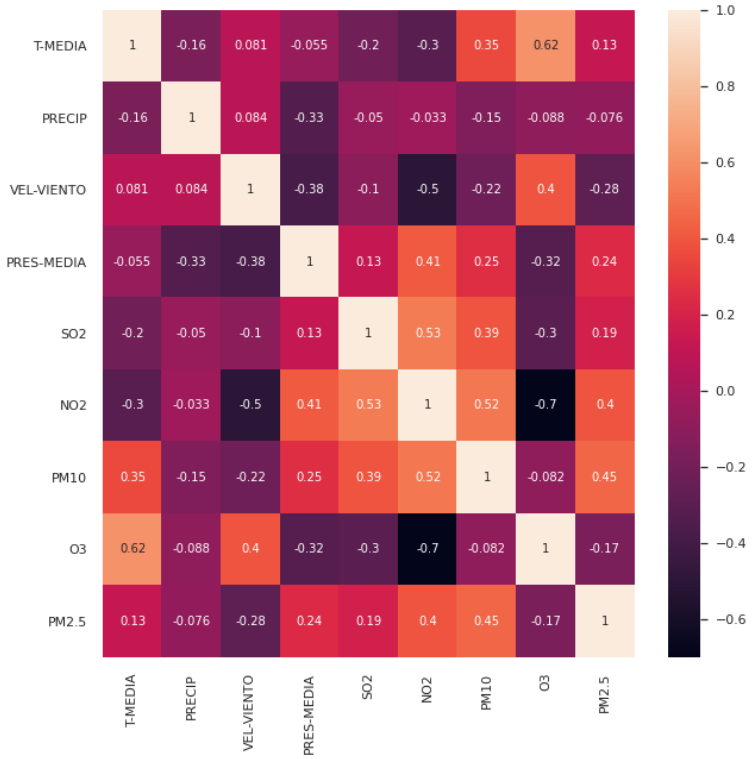


Figura 40. Matriz de correlaciones entre las distintas magnitudes

Las correlaciones tienen un máximo de 1 y un mínimo de -1. Comprobando la Figura 40, las magnitudes que más fuertemente relacionadas están son el NO2 y el O3 con un coeficiente de -0.7. Este coeficiente indica que cuando una de las dos magnitudes aumenta, la otra disminuye en cierta medida.

Figura 41. Comprobación correlación lineal entre NO2 y O3

Utilizando los datos del año 2017 como muestra para comprobar cómo se afectan mutuamente, se observa en la Figura 41 como se cumple esta relación lineal.

4.3.2. Modelo clásico de regresión

Los modelos clásicos de regresión son modelos estadísticos que se ajustan a datos pasados y presentes para intentar predecir los valores futuros. El más extendido es el Modelo Autoregresivo Integrado de Media Móvil (*Autoregressive Integrated Moving Average* o ARIMA) aunque tiene un amplio número de variantes en función de las características de los datos.

Los principales algoritmos relacionados con ARIMA son:

- ARIMA: utilizado principalmente para la regresión de datos que siguen una tendencia, pero sin componente estacional. En función de la configuración de los parámetros, posibilita la implementación del algoritmo AR (autorregresivo o *Autoregressive*), MA (*Moving Average*) o ARMA (AR+MA).

- SARIMA: variante del modelo ARIMA enfocado a la regresión de datos que siguen una tendencia con componente estacional. En función de la configuración de los parámetros, posibilita la implementación de AR, MA, ARMA o ARIMA.

- SARIMAX: variante del modelo SARIMA que incluye variables externas al modelo, también llamadas exógenas. En función de la configuración de sus parámetros permite la implementación de los mismos modelos que SARIMA pudiendo añadir variables exógenas.

La elección del modelo de regresión se hace en función de la estacionalidad de los datos de entrada.

El modelo ARIMA consta de 3 parámetros de configuración, cada uno asociado a:

- p: número de datos retrasados incluidos en la observación del modelo autorregresivo (AR).
- d: diferencia de una observación sin procesar con una anterior. Este parámetro es utilizado para convertir los datos en estacionarios.
- q: número de observaciones incluidas en el modelo de Media Móvil (MA).

La predicción utilizando el modelo clásico se hará sobre los datos de dióxido de nitrógeno del año 2019, dado que el conjunto de datos de todos los años es demasiado para procesarlo con dichos algoritmos en el entorno de desarrollo utilizado.

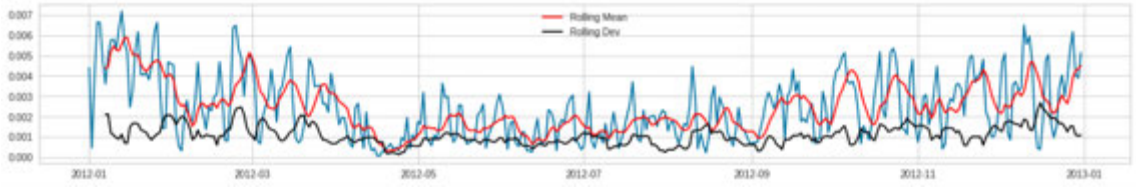


Figura 42. Datos de NO2 del año 2012 (línea azul) junto con la media semanal móvil (línea roja) y la desviación típica semanal (línea negra)

El primer paso para analizar los datos que van a ser utilizados es comprobar si son estacionales. Para realizar esta comprobación se ha utilizado la prueba de Dickey-Fuller para el contraste de hipótesis.

Prueba de Dickey-Fuller aumentada

La prueba de Dickey-Fuller aumentada (ADF) realiza un contraste de hipótesis para detectar si una serie de tiempo tiene una unidad raíz, es decir, está definido por una tendencia. Se formulan dos hipótesis:

- H0 (hipótesis nula): los datos tienen unidad raíz por lo que no son estacionales.
- H1 (hipótesis alternativa): los datos no tienen unidad raíz así que son estacionales.

La prueba genera un p-valor que es usado para rechazar la hipótesis nula si tiene un valor superior a 0,05. En el caso de los datos de NO2 en 2019, el resultado obtenido es:

Resultado	Valor
p-valor	0.12159836736332252

Tabla 14. Resultado Dickey-Fuller sobre los dados del año 2012 de NO2

Dado que los datos de NO2 no son estacionales por ser el p-valor > 0.05, se va a utilizar el algoritmo ARIMA.

Los parámetros del modelo ARIMA pueden ser estimados manualmente utilizando la autocorrelación y correlación parcial de los datos de NO2 a través del número de retrasos. Sin embargo, con el objetivo de hallar los valores más adecuados, se hará uso de una función para probar un rango de valores de cada parámetro y elegir el que mejor se ajuste a los datos.

El conjunto de los datos se ha dividido en dos grupos:

- Datos de entrenamiento: conjunto de datos con a los que ajustar el modelo ARIMA.
- Datos de prueba: conjunto de datos sobre los que se compararan las predicciones y se calcula el error.

Cabe destacar que no se va a hacer una sola predicción de tantos valores como valores de prueba haya, sino que se van a realizar tantas predicciones del siguiente valor como datos de prueba. Es decir, no se harán una predicción de X valores futuros, sino que se harán X predicciones de un único valor. Cada vez que se haga una predicción, se volverá a ajustar el modelo con los datos de entrenamiento más el valor de prueba correspondiente. El objetivo de este modo de predicción es reducir el error lo máximo posible.

Los parámetros del modelo ARIMA que mejor se ajustan a los datos de entrenamiento son p=8, d=1 y q=0. Esto indica que el modelo de la Media Móvil (*MA*) no influye en el resultado de la predicción.

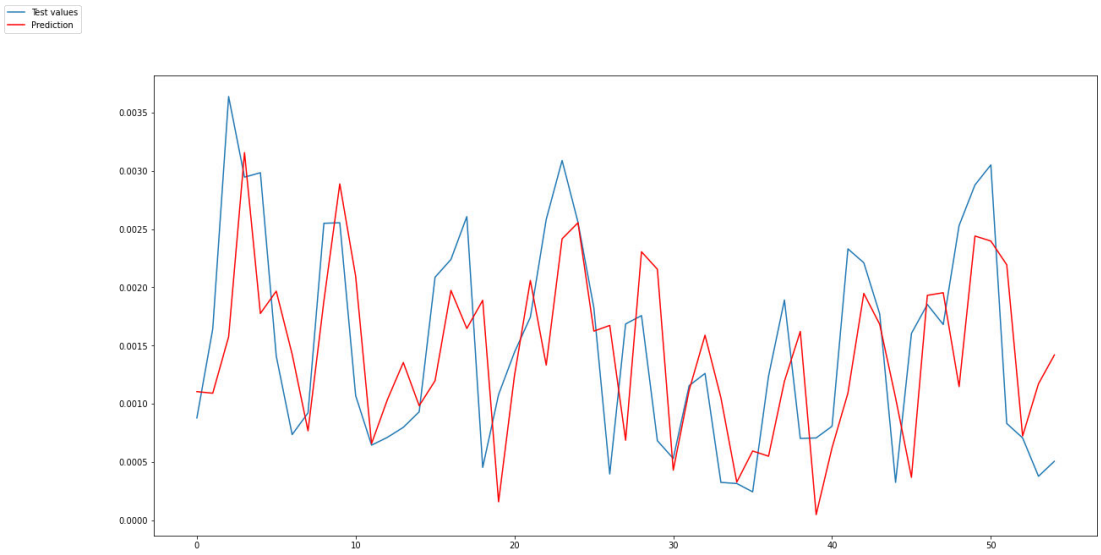


Figura 43. Predicciones diarias del modelo ARIMA

4.3.3. LSTM

Se han utilizado distintas arquitecturas neuronales para comparar los resultados de los distintos modelos a la hora de predecir futuros valores de las medidas de contaminación.

Las predicciones se han hecho sobre los datos temporales con muestreo horario, aunque a la hora de visualizarlas se ha optado por una agrupación de datos semanal para poder distinguir visualmente la diferencia entre el valor real y la predicción.

Para el entrenamiento de cada una de las redes neuronales se han especificado los mismos parámetros (llamados hiperparámetros) y, de esta forma, poder comparar el rendimiento:

Parámetro	Valor
Batch-size	8
Epoch	100
Optimizador	Adam
Función de pérdida	Error medio cuadrático (MSE)

Tabla 15. Hiperparámetros definidos

LSTM simple (Vanilla LSTM)

Una red neuronal Vanilla es una red neuronal con una arquitectura muy simple, utilizada simplemente para ilustrar cuál es el funcionamiento o una muestra de los resultados.

Para este objetivo, se ha utilizado únicamente una capa oculta 8 neuronas LSTMs y una capa de salida con una neurona totalmente conectada a la capa anterior que genera el próximo valor de la magnitud que se está estudiando.

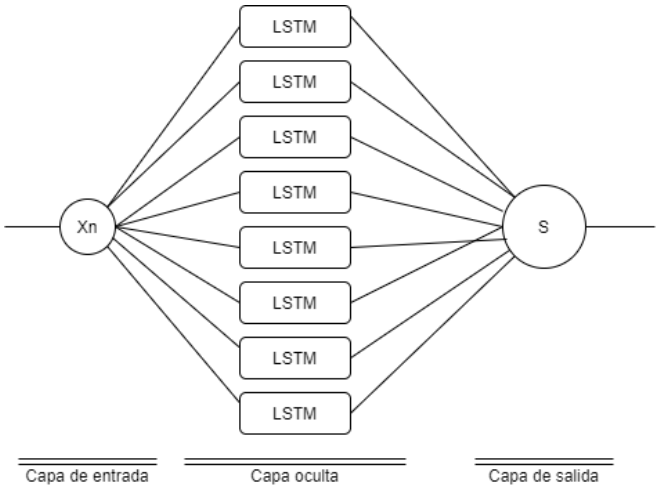


Figura 44. Arquitectura red neuronal simple

Esta red neuronal se ha entrenado utilizando, por un lado, predicción univariante con los datos de NO2 y, por otro lado, predicción multivariante de los futuros datos de NO2 a partir de todas las magnitudes disponibles.

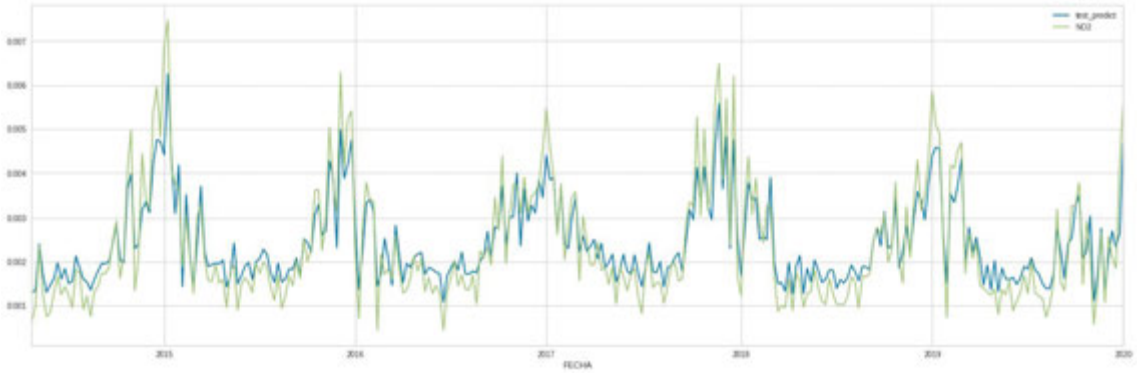


Figura 45. Predicciones semanales de la red neuronal simple univariante

Figura 46. Predicciones semanales de la red neuronal simple multivariante

A simple vista no parece que haya mucha diferencia entre los resultados de la Figura 45 y la Figura 46, aunque los resultados del error de las predicciones se comentarán más adelante.

Puede compararse la función de perdida de cada una de las redes neuronales a través de las *epochs* para comprobar cómo se ha ido adaptando el modelo a través del entrenamiento en base a la función de pérdida. En la Figura 47 y la Figura 48 se representan la función de pérdida de los datos de entrenamiento (color azul) y de los datos de validación (color verde).

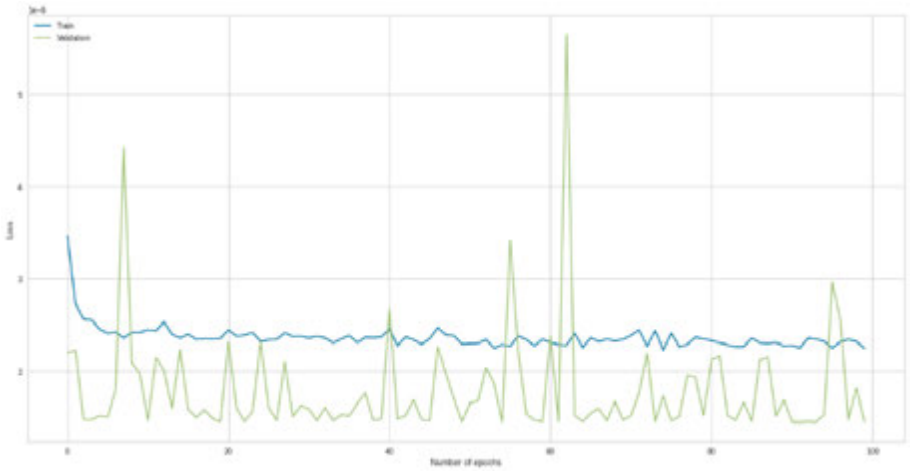


Figura 47. Función de perdida de la red neuronal simple univariante

En la Figura 47 se ve que el error es muy reducido ya en la primera *epoch* y continúa descendiendo y estabilizándose mientras que el error en la Figura 48 necesita más epochs para estabilizarse, aunque logra converger con el error de validación a lo largo de todo el entrenamiento.

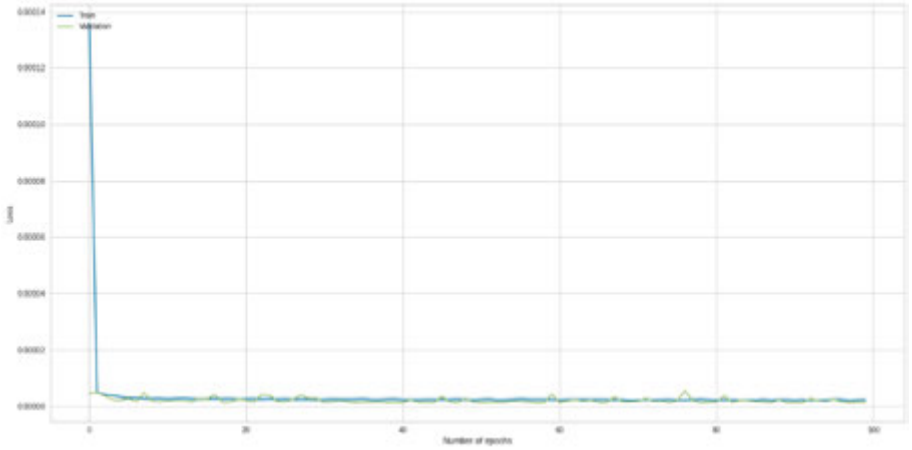


Figura 48. Función de perdida de la red neuronal simple multivariante

LSTM multicapa

Con el objetivo de comparar los resultados de la red neuronal simple con una red neuronal más compleja, se ha aplicado el mismo proceso de entrenamiento y predicción a otra red neuronal.

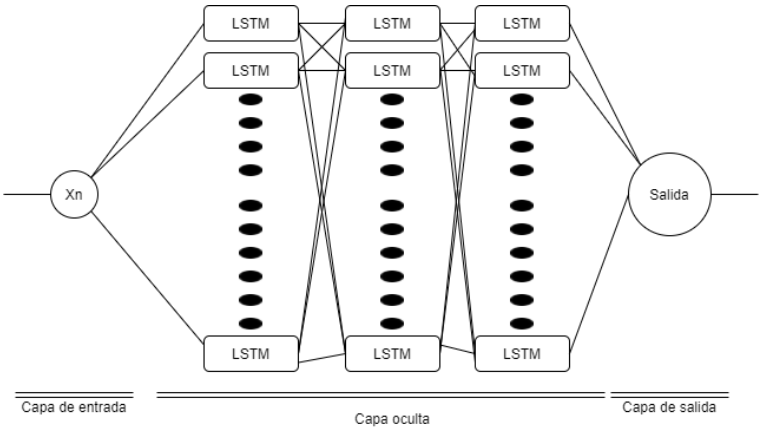


Figura 49. Arquitectura red neuronal multicapa

La red neuronal en cuestión consta de 3 capas ocultas, cada una con 8 neuronas, una capa de entrada y otra de salida, como se puede comprobar en la Figura 49.

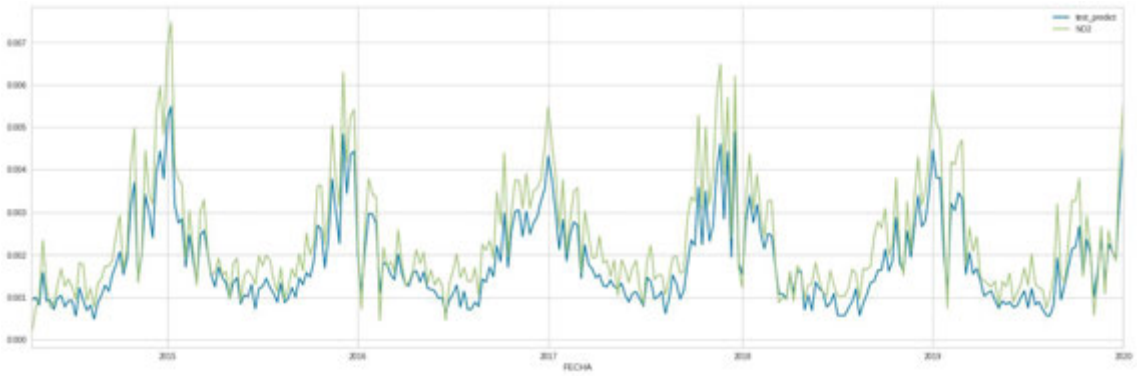


Figura 50. Predicciones semanales de la red neuronal multicapa multivariante

Las predicciones parecen variar mínimamente en comparación con la red neuronal *Vanilla* mientras que la función de pérdida, según la Figura 50, parece estabilizarse a partir de la *epoch* 30.

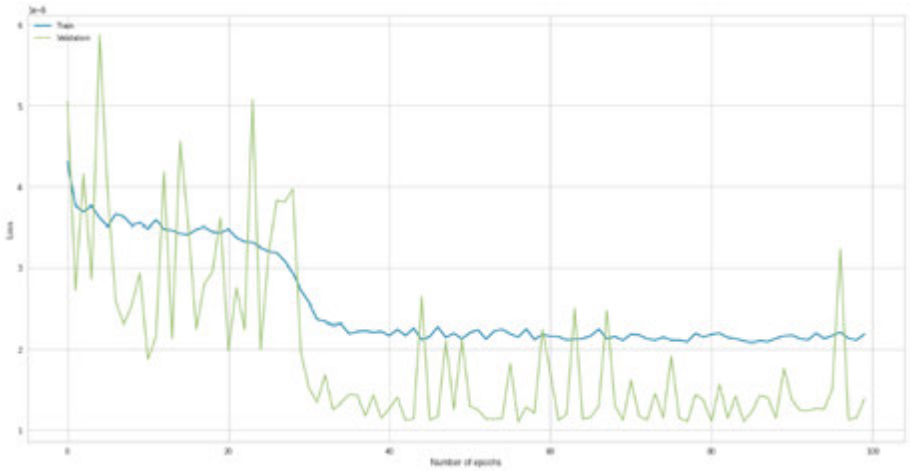


Figura 51. Función de perdida de la red neuronal multicapa multivariante

4.4. Clustering

A continuación, se van a explicar el procedimiento seguido para implementar los algoritmos de clustering y dos alternativas a la hora de visualizar los resultados en un gráfico, así como una valoración de los resultados.

Estos algoritmos de clustering son aplicados sobre los datos completos de contaminación y meteorológicos, con el formato de la Tabla 11.

4.4.1. Pasos Previos

Como se menciona en el apartado 2.4.1.1, se debe especificar el número de grupos en los que se van a dividir las observaciones por lo que la pregunta que hay que hacerse es ¿qué número de grupos es el idóneo para mis datos?

Para responder a esta pregunta se han utilizado tres métodos:

- *Método del codo (Elbow Method)*

El Método del Codo es una técnica muy popular para calcular el número óptimo de grupos en los que dividir un cierto conjunto de datos. Para ello se le indica un rango de posibles números de grupos y, por cada uno, calcula la suma de las distancias al cuadrado de cada punto a su hipotético correspondiente grupo [29].

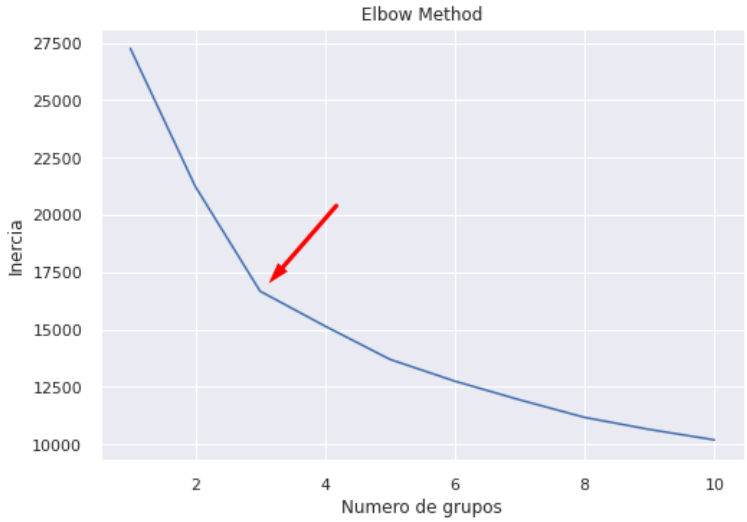


Figura 52. Resultado del Método del Codo (*Elbow Method*)

A esta técnica se le llama método del codo porque el número óptimo de clústeres se da en el punto en que la curva se aplana y parece el codo de un brazo (punto indicado por la fecha en la Figura 52), por lo que el resultado son 3 grupos.

- *Silhouette*

Es una medida de cuán similar es un objeto a su propio grupo en comparación con el resto [30]. El resultado es el máximo valor de la media de todos los datos con un rango de (-1, 1) y, cuanto más alto es, más similares son los datos a su grupo y más distinto de los grupos vecinos.

K (Número de Clústeres)	**Silhoutte**
2	0.2502
3	0.2638
4	0.2210
5	0.2144
6	0.1869

Tabla 16. Silhoutte Score para distintos números de grupos o clústeres

Al revisar los resultados de la puntuación Silhoutte en la Tabla 16, el resultado con mayor puntuación es utilizar 3 grupos o clústeres, que coincide con el resultado del Método del Codo por lo que es definitivamente el número de clústeres sobre el que se va a aplicar el algoritmo de K-Medias.

4.4.2. K-Medias (o K-Means)

El algoritmo de K-Medias es un algoritmo de aprendizaje no supervisado muy utilizado para la partición de observaciones en k grupos en función de su similitud. Utilizando la librería de Python *scikit-learn* su implementación es muy sencilla y rápida peor hay que tener en cuenta unos requisitos previos.

Una vez aplicado el algoritmo, a cada observación se le asigna uno de los 3 clústeres (identificados con los valores 0, 1 y 2) en una nueva columna del marco de datos.

Fecha	T-Media	Precip	Vel-Viento	Pres-Media	SO2	NO2	PM10	O3	PM2.5	Escenario	K-Means
2001-01-01 01:00:00	-0.964	1.116	0.262	-1.097	0.295	0.254	0.113	-0.551	-0.296	0	0

Tabla 17. Marco de datos tras la aplicación del algoritmo K-Medias.

4.4.3. Clústering Jerárquico

De igual manera que el algoritmo K-Medias, el Clústering Jerárquico es una técnica muy popular de aprendizaje no supervisado para el agrupamiento de datos no estructurados.

El resultado de este algoritmo es un dendograma con la distancia euclídea (es decir, su similitud) entre los distintos puntos y donde puede verse cómo se van juntando los datos en grupos cada vez más grandes hasta que todos han sido abarcados. El número de clústeres óptimo se elige por criterio subjetivo, al igual que el algoritmo K-Means.

En el dendograma se puede marcar un "límite" de distancia Euclídea y línea marca el número de clústeres.

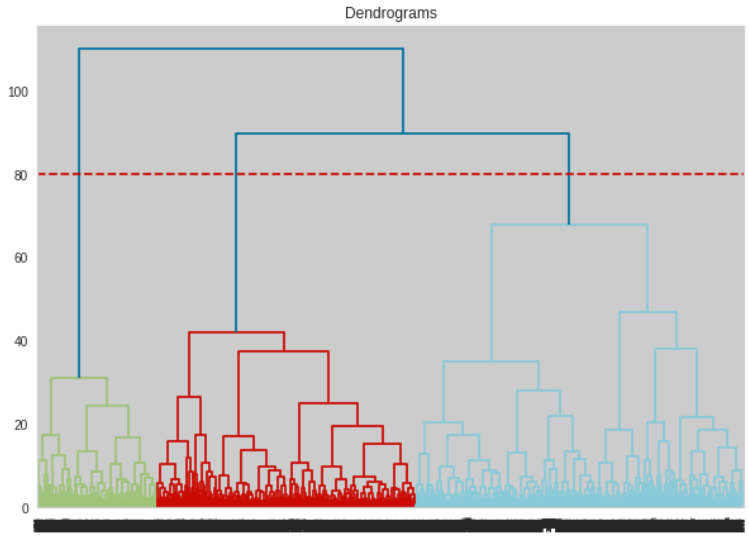


Figura 53. Dendograma con un límite de 3 clústeres

Siguiendo el criterio marcado por el Método del Codo, la puntuación Silhoutte y haciendo un análisis visual al dendograma, una posible solución es marcar el límite en 80 como en

la Figura 53, lo que deja el número de clústeres en 3. Con esta solución se aplica el mismo procedimiento que K-Medias creándose una nueva columna en el marco de datos con el grupo al que pertenece cada una de las observaciones.

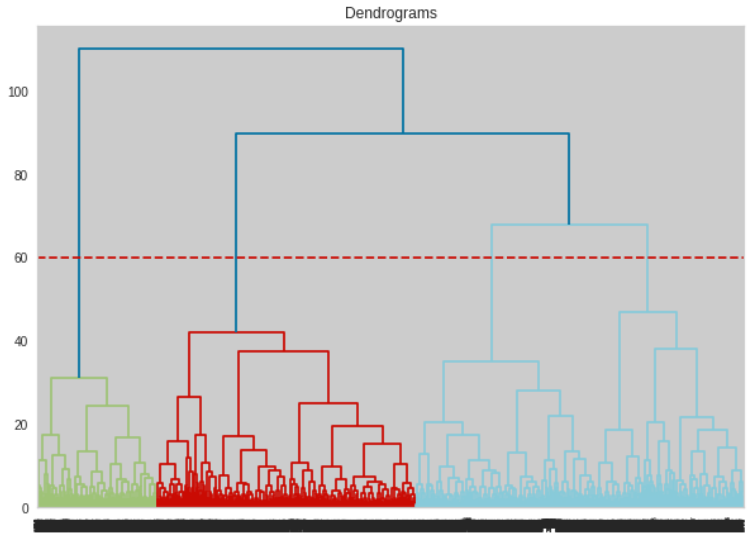


Figura 54. Dendograma con un límite de 4 clústeres

Otra posible solución puede ser marcar el límite en 60, como en la Figura 54, lo que dejaría 4 clústeres en los que agrupar los datos. Esto puede ser una posible solución ya que no se aleja mucho del límite marcado previamente por lo que se crea otra columna en el marco de datos, pero esta vez con la pertenencia de cada observación a uno de los 4 grupos.

Por defecto se han distinguido tres clústeres principales (color verde, color rojo y color azul), lo que concuerda con el criterio elegido para el algoritmo K-Medias. Sin embargo, esta representación en formato de dendograma muestra que el tercer clúster (color azul) está compuesto por dos grupos que se hayan a una distancia considerable entre sí. Por ello, el criterio aplicado en la Figura 53 separa los datos en grupos que se hayan aproximadamente a la misma distancia, es decir, con la misma similitud entre ellos.

5. RESULTADOS OBTENIDOS

A continuación, se van a comentar los resultados obtenidos en los distintos apartados del capítulo anterior dividiéndose en secciones según correspondan.

5.1. Análisis inicial de los datos

El análisis de los datos históricos es interpretable, más que de una forma tecnológica, desde un punto de vista centrado en los problemas presentados en la sección 1.1 pues es ahí donde se pone en contexto el papel de los contaminantes atmosféricos.

Las condiciones atmosféricas se mantienen estables a través de todo el marco de tiempo contemplado en este estudio, si bien sería necesario realizar un estudio mucho más amplio en el tiempo para sacar conclusiones fiables.

Por otro lado, una gran mayoría de los contaminantes han visto sus niveles reducidos en los últimos años, en especial el SO2 y NO2 han sufrido una reducción muy pronunciada a partir del año 2010. Sin embargo, los niveles de O3 han aumentado drásticamente, lo que puede afectar gravemente a la salud de las personas [31].

5.2. Regresión

El enfoque principal para predecir los valores futuros ha sido aplicar Deep Learning, variando el diseño de la red neuronal, pero manteniendo los hiperparámetros para poder comparar los resultados con mayor fiabilidad. Se ha utilizado el modelo ARIMA para comparar los resultados entre los métodos clásicos de predicción y los actuales algoritmos de aprendizaje automático, si bien los datos sobre los que se ha calculado el error son un subconjunto de los datos utilizados con las redes neuronales.

Modelo	**Error (MSE)**
ARIMA	6.10229770762e-07
LSTM simple univariante	2.27970209444e-06
LSTM simple multivariante	5.78301387732e-07
LSTM multicapa multivariante	1.19993757717e-06

Tabla 18. Error de los modelos de regresión

Es interesante comprobar en la Tabla 18 cómo el error del modelo más complejo de todos (LSTM multicapa multivariante) no es el más bajo de todos. Esto se debe a que una red neuronal más extensa no implica necesariamente que su funcionamiento vaya a ser mejor, ya que el diseño del modelo debe ajustarse a las características de los datos implicados y es necesario realizar múltiples experimentos con distintas combinaciones de redes neuronales e hiperparámetros.

El error del modelo ARIMA es menor que el de algunos modelos de Deep Learning, aunque hay que recordar que los datos sobre los que se ha ajustado el modelo son los correspondientes a un solo año. Sin embargo, esto demuestra que las técnicas clásicas de regresión pueden competir en muchas ocasiones con algoritmos más actuales y complejos y no deben ser descartadas a la hora de resolver problemas.

5.3. Clustering

Aunque todas las observaciones han sido agrupadas en su correspondiente clúster por varios algoritmos, es recomendable buscar alguna manera de representar gráficamente estas agrupaciones para tener una visión general.

Fecha	T-Media	Precip	Vel-Viento	Pres-Media	SO2	NO2	PM10	O3	PM2.5	Escenario	K-Means	Hierach_3	Hirach_4
2003-01-01 01:00:00	-0.96	1.11	0.26	-1.09	0.29	0.25	0.11	-0.55	-0.29	0	1	0	0

Tabla 19. Ejemplo de datos tras aplicar algoritmos de clustering.

Sin embargo, este agrupamiento se ha hecho con la representación de los datos obtenidos de temperatura, precipitaciones, velocidad del viento, presión atmosférica, SO2, NO2, PM10, O3 y PM2.5, es decir que están compuestos por 9 dimensiones y resulta imposible representarlos gráficamente.

En este punto hay dos posibles opciones para estudiar los resultados del clustering: estudiar los clústeres en cada una de las variables o aplicar técnicas de reducción de dimensionalidad para representar toda la información posible en un espacio de dos dimensiones.

5.3.1. Estudio de cada una de las magnitudes

Realizando el estudio de cada magnitud por separado se pretende identificar patrones en los clústeres asociados a los datos a través de cada una de las variables.

El primer paso es comprobar los datos históricos distinguiendo los distintos clústeres de cada una de las magnitudes. En la Figura 55 se muestran los datos de cada una de las

variables a lo largo del tiempo con diferenciando el clúster asociado a cada punto con un color distinto.

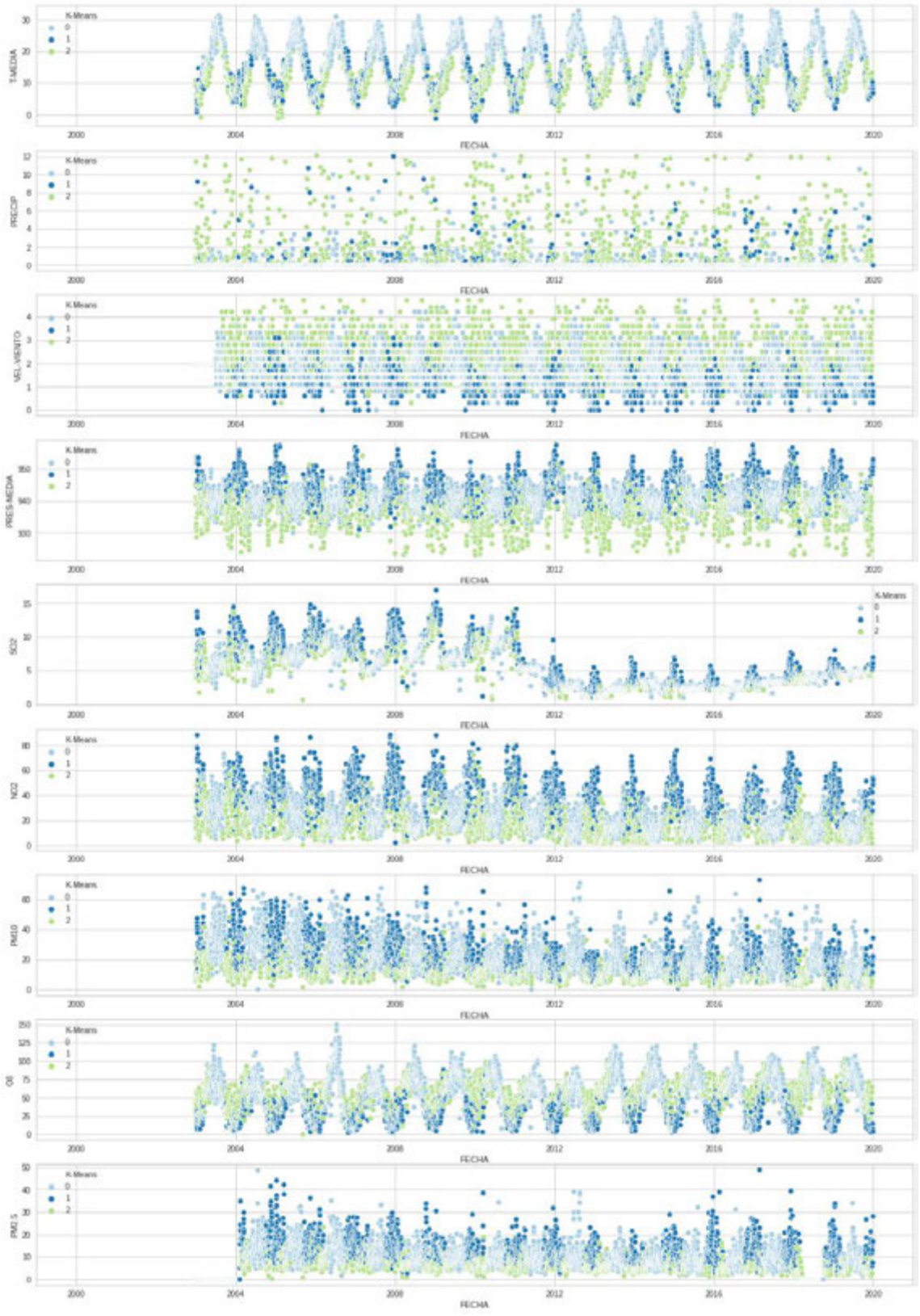


Figura 55. Visión particular de los resultados de K-Medias

Los resultados de los 3 algoritmos de clustering siguen un comportamiento muy similar a la hora de clasificar los datos por lo que se van a utilizar los resultados de K-Medias para ilustrar los criterios ya que parecen bastante generalizados.

La primera conclusión a la que se llega observando la Figura 55 es que los criterios se repiten cada año por lo que es redundante observar todo el conjunto. Un posible enfoque puede ser visualizar datos en un año en concreto (2019, por ejemplo) y marcando los finales de cada estación (en orden son final de invierno, final de primavera, final de otoño e invierno de nuevo).

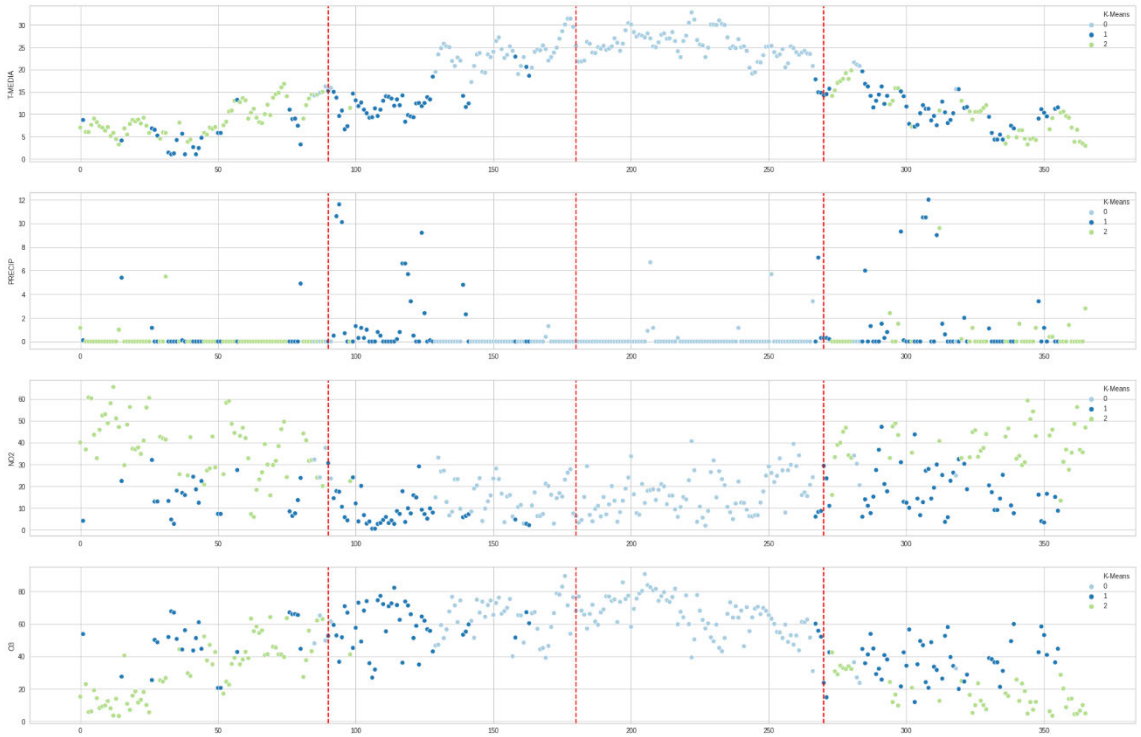


Figura 56. Resultados K-Medias diferenciando estaciones del año

Salta a la vista en la Figura 56 que hay un componente temporal muy marcado a la hora de clasificar los datos en los distintos clústeres. Por ejemplo, los datos de verano pertenecen (casi en su mayoría) al clúster 0 y los de invierno al clúster 2, repartiéndose los de otoño y primavera en el clúster 1.

Este criterio estacional parece haber sido el criterio principal a la hora de agrupar los datos, lo cual tiene sentido considerando que las condiciones meteorológicas están muy marcadas por las estaciones e influyen directamente sobre los contaminantes.

5.3.2. Reducción de dimensionalidad

Existen técnicas para reducir el número de dimensiones de los datos a la vez que se mantiene la información de los datos como el Análisis de Principales Componentes (o *Principal Componente Analysis* o PCA) y t-SNE (*t-distributed Stochastic Neighbor Embedding*).

El objetivo de representar los datos utilizando este tipo de algoritmos es evaluar cómo se agrupan los datos por similitud y, en base a esto, evaluar cómo han actuado los algoritmos de clustering y si se diferenciar los grupos claramente.

Tras aplicar cada uno de los algoritmos, se disponen de dos dimensiones para cada uno que podríamos llamar la coordenada X y la coordenada Y. Utilizando estas coordenadas se han representado los datos en cuatro gráficos:

1. Pertenencia de los datos divididos en grupos por el algoritmo K-Medias.

2. Identificación de los escenarios anticontaminación. Se ha decidido representar esta categoría por si se diera el caso de que algún algoritmo de agrupamiento ha separado los escenarios más graves en un mismo grupo o clúster.

3. Pertenencia de los datos divididos en tres grupos por el agrupamiento jerárquico.

4. Pertenencia de los datos divididos en cuatro grupos por el agrupamiento jerárquico.

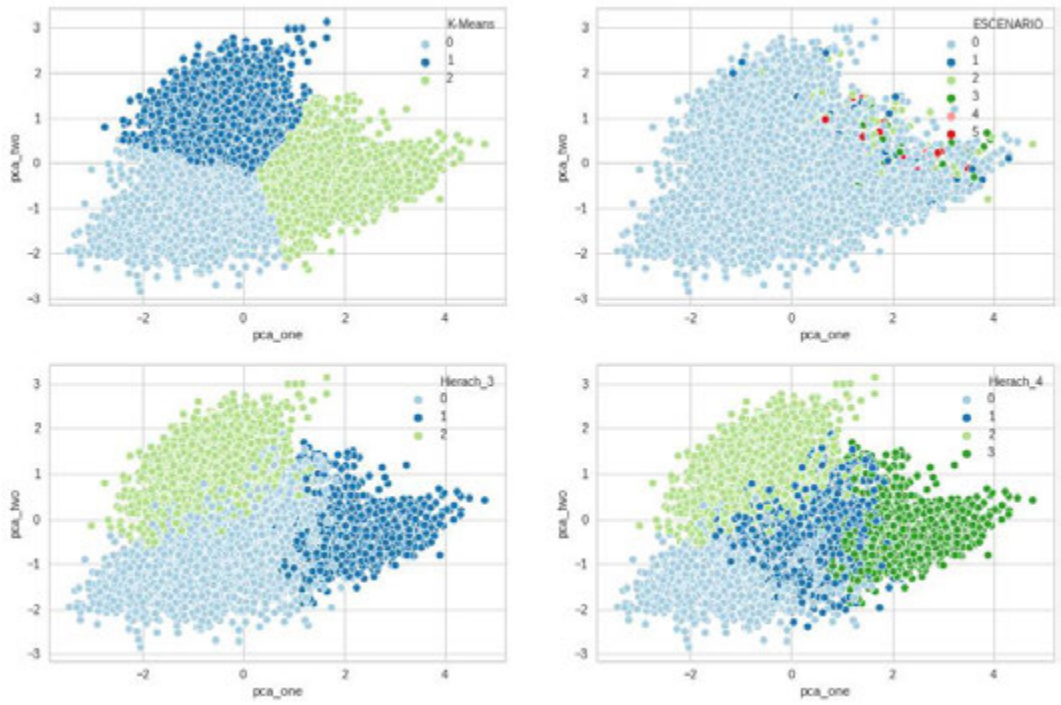


Figura 57. Representación en 2D de los resultados del clustering utilizando PCA

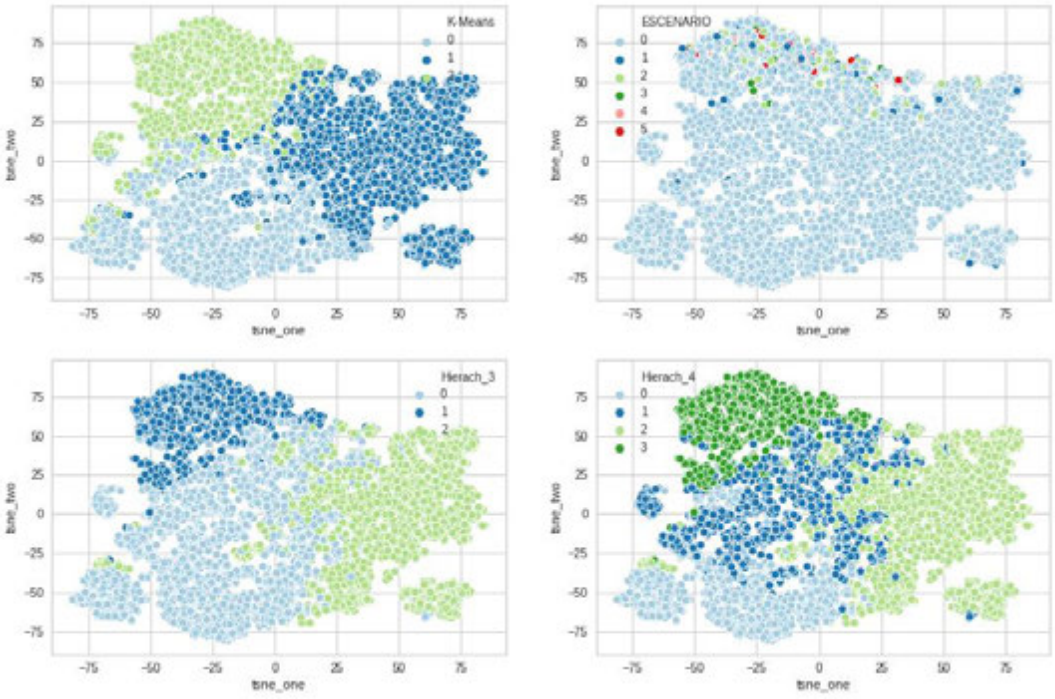


Figura 58. Representación en 2D de los resultados del clustering utilizando t-SNE

Realizando una comparación visual entre los resultados del PCA y t-SNE (Figura 57 y Figura 58, respectivamente) se puede comprobar que con t-SNE, por normal general, los datos más similares se agrupan entre sí mientras que se separan del resto, en definitiva, quedan grupos más diferenciados.

Aunque ambos algoritmos cumplen el objetivo de representar los datos en un espacio dimensional reducido, queda patente que t-SNE es ampliamente utilizado en el análisis y ciencia de datos para la visualización de datos en un espacio dimensional de 2 o 3 dimensiones [32].

Otro argumento en contra del PCA es que la cantidad de información que puede representar varía en función del número de componentes elegidos.

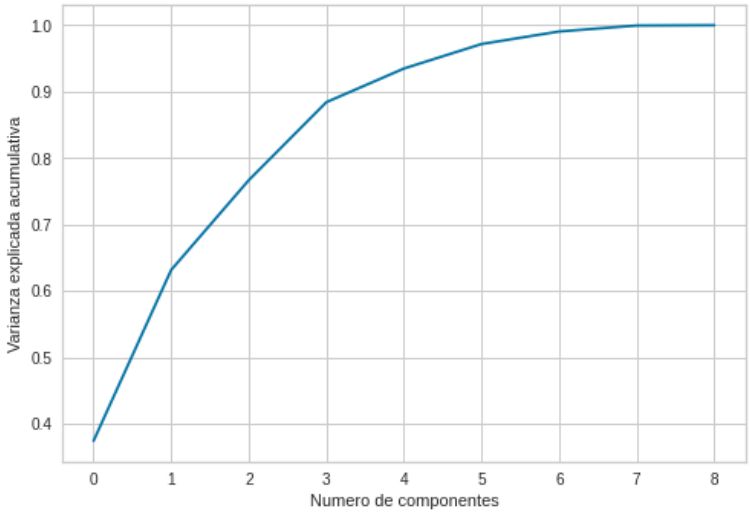


Figura 59. Varianza explicada acumulativa

En la Figura 59 se muestra que, con los datos que disponemos, con 2 componentes se representa aproximadamente el 77% de los datos, lo cual no es una buena cifra ya que suele ser aceptable un 90% o más por lo que la información representada con los resultados de PCA no debe ser considerada como representativa, sino simplemente ilustrativa.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. Conclusiones

En vista de los resultados obtenidos y partiendo de los objetivos definidos al comienzo del presente documento, se han obtenido las siguientes conclusiones:

I. Los datos disponibles para realizar un estudio de las condiciones meteorológicas y de contaminación atmosférica son accesibles para cualquier persona ya que los requisitos previos son mínimos:

Requisito	Razón
Conocimientos de peticiones HTTP	Los datos meteorológicos son accesibles por peticiones HTTP a través de peticiones la API REST de AEMET.
Correo electrónico	Necesario para recibir la clave de la API de AEMET para solicitar los datos atmosféricos por peticiones HTTP.
Acceso a Internet	Indispensable acceso a la red para poder descargar los datos de contaminación atmosférica en la página del Ayuntamiento de Madrid y para obtener los datos meteorológicos.

Tabla 20. Requisitos para obtener los datos

Los datos obtenidos son suficientes para realizar el objetivo de este proyecto y se han ilustrado los pasos necesarios para unificar los datos originales en un formato más adecuado para su estudio.

II. En el transcurso de este proyecto, prácticamente la totalidad de las modificaciones que han sufrido los datos y su visualización se han llevado a cabo a través de la librería *pandas*, una de las herramientas gratuitas de análisis de datos más utilizadas en la actualidad.

Software	2019 % share	2018 % share	2017 % share
Python	65.8%	65.6%	59.0%
RapidMiner	51.2%	52.7%	31.9%
R Language	46.6%	48.5%	56.6%
Excel	34.8%	39.1%	31.5%
Anaconda	33.9%	33.4%	24.3%
SQL Language	32.8%	39.6%	39.2%
Tensorflow	31.7%	29.9%	22.7%
Keras	26.6%	22.2%	10.7%
scikit-learn	25.5%	24.4%	21.9%
Tableau	22.1%	26.4%	21.8%
Apache Spark	21.0%	21.5%	25.5%

Tabla 21. Software más utilizado de data Science/Análisis/Machine Learning [33]

Además de *pandas*, a lo largo del proyecto se han utilizado librerías que están entre las más utilizadas como Keras para el diseño de la red neuronal y scikit-learn en la implementación de algoritmos de aprendizaje no supervisado.

III. Se han analizado y visualizado los datos históricos de cada una de las magnitudes, tanto meteorológicas como de contaminación atmosférica a través de gráficos con toda la línea de tiempo y, en los casos en los que ha sido necesario, se ha especificado un rango de tiempo específico.

Por ejemplo, la magnitud de PM2.5 solo empezó a registrarse a partir del año 2005 por lo que merecía un trato específico sobre el resto.

En el caso de que los datos mostraran un comportamiento estacional, se han utilizado las herramientas necesarias para descomponerlos en los componentes estacionales (datos observados, tendencia, estacionalidad, residual) con el fin de estudiar la tendencia a lo largo del periodo de tiempo definido.

Por último, se ha demostrado la utilidad y las implicaciones del coeficiente de correlación lineal a la hora de analizar las relaciones lineales entre las distintas magnitudes.

IV. Gracias al desarrollo del proyecto se ha logrado un considerable manejo de las herramientas utilizadas y en el lenguaje de programación Python.

Es destacable la interacción que puede se descubre entre las distintas librerías. Por ejemplo, funciones de *pandas* para visualizar datos de sus propios marcos de datos utilizan librerías propias de visualización como backend para lograr un estilo más descriptivo, o *numpy*, una librería muy extensa de funciones matemáticas puede utilizarse para preparar los datos de entrada a una red neural o manipular los datos de *pandas*.

También se ha llegado a comprender los mecanismos de Python para la estructura del código a la hora de lograr que sea ilustrativo, utilizando comentarios y nombres de funciones descriptivos siguiendo las mejores prácticas [34].

V. Se han utilizado algoritmos de clustering con el objetivo de identificar patrones en los datos disponibles que permitan definir escenarios de contaminación, ampliando así los ya existentes en la ciudad de Madrid.

No se ha logrado dicho objetivo puesto que el resultado de estos algoritmos ha resultado ser la agrupación de los datos en base a la estación del año a la que pertenecen, en su mayor o menor medida. Esto pone de manifiesto que, si bien no es el resultado deseado, los algoritmos han identificado las similitudes entre los datos satisfactoriamente, pues se ha podido comprobar que gran parte de las magnitudes tienen un comportamiento común a lo largo de cada año.

VI. Se han utilizado varios algoritmos y técnicas de Machine Learning con distintas aplicaciones y objetivos. Para cada uno de ellos se ha explicado, con palabras y gráficamente, el objetivo de su implementación y los resultados que se han obtenido.

Los resultados de los algoritmos de aprendizaje no supervisado, al no poder medir la calidad de una forma objetiva, han sido analizados de distintas formas y evaluados de forma subjetivo y de acuerdo con los objetivos marcados en la realización de este proyecto. Los algoritmos de aprendizaje supervisado han sido comparados entre ellos utilizando el error de sus predicciones con los datos reales y su proceso de aprendizaje explicado para entender cómo han llegado a esas conclusiones.

VII. Los algoritmos de regresión han resultado ser muy precisos a la hora de generar los futuros valores de los datos. Al ser datos viables y sobre el contaminante que determina los escenarios de contaminación, es posible predecir los datos de cada

una de las estaciones de la ciudad (pues es necesario tener en cuenta las distintas estaciones de las distintas áreas urbanas) para calcular los criterios que se aplicarían en esos supuestos datos futuros y, de esta forma, actuar de forma proactiva antes de llegar a niveles límite de contaminación si fuera necesario.

6.2. Trabajos futuros

Los trabajos futuros necesarios para ampliar la información del presente estudio pueden separarse en varios apartados:

- Ampliación de los datos utilizados con datos de población en la ciudad de Madrid, datos del tráfico en las horas a las que se hicieron las mediciones de contaminantes y gasto energético asociado a distintos combustibles.

- Aplicar técnicas de clustering más avanzadas que permitan la minería de datos teniendo en cuenta el componente secuencial de los datos y no simplemente un como muestras de datos independientes. Algunos de estos algoritmos son Generalized Sequential Pattern (GSP) [38], Sequential PAttern Discovery using Equivalence classes (SPADE) [39] y Sequential PAttern Mining (SPAM) [40].

- Definir un criterio de parametrización y arquitectura de las redes neuronales para maximizar la calidad de las predicciones.

- Implementar redes neuronales más complejas y actuales que las LSTMs como la GRU o *Transformer*, también orientadas a la predicción de series temporales.

- Con el objetivo de ampliar los conocimientos de Python, crear una *dashboard* alojado en la web donde se puedan consultar las conclusiones de este estudio utilizando las librerías Dash y Plotly para la creación de gráficos interactivos.

- Ampliar el estudio con la relación entre las distintas estaciones de medición, posiblemente basándose en la distancia y la situación de cada una.

7. PLANIFICACIÓN Y PRESUPUESTO

En el presento anexo se detalla una aproximación del tiempo dedicado a la elaboración de este trabajo, junto con un desglose de los costes asociados a la realización de este.

1. Planificación del proyecto

La totalidad de la realización del proyecto se ha dividido en cada una de los capítulos descritos en este documento. Una forma muy extendida de ilustrar el tiempo dedicado a cada sección es el diagrama de Gantt, que se basa en el número de horas dedicada a cada sección.

La planificación se ha realizado teniendo en cuenta la jornada ordinada laboral, con 8 horas diarias de trabajo y de lunes a viernes.

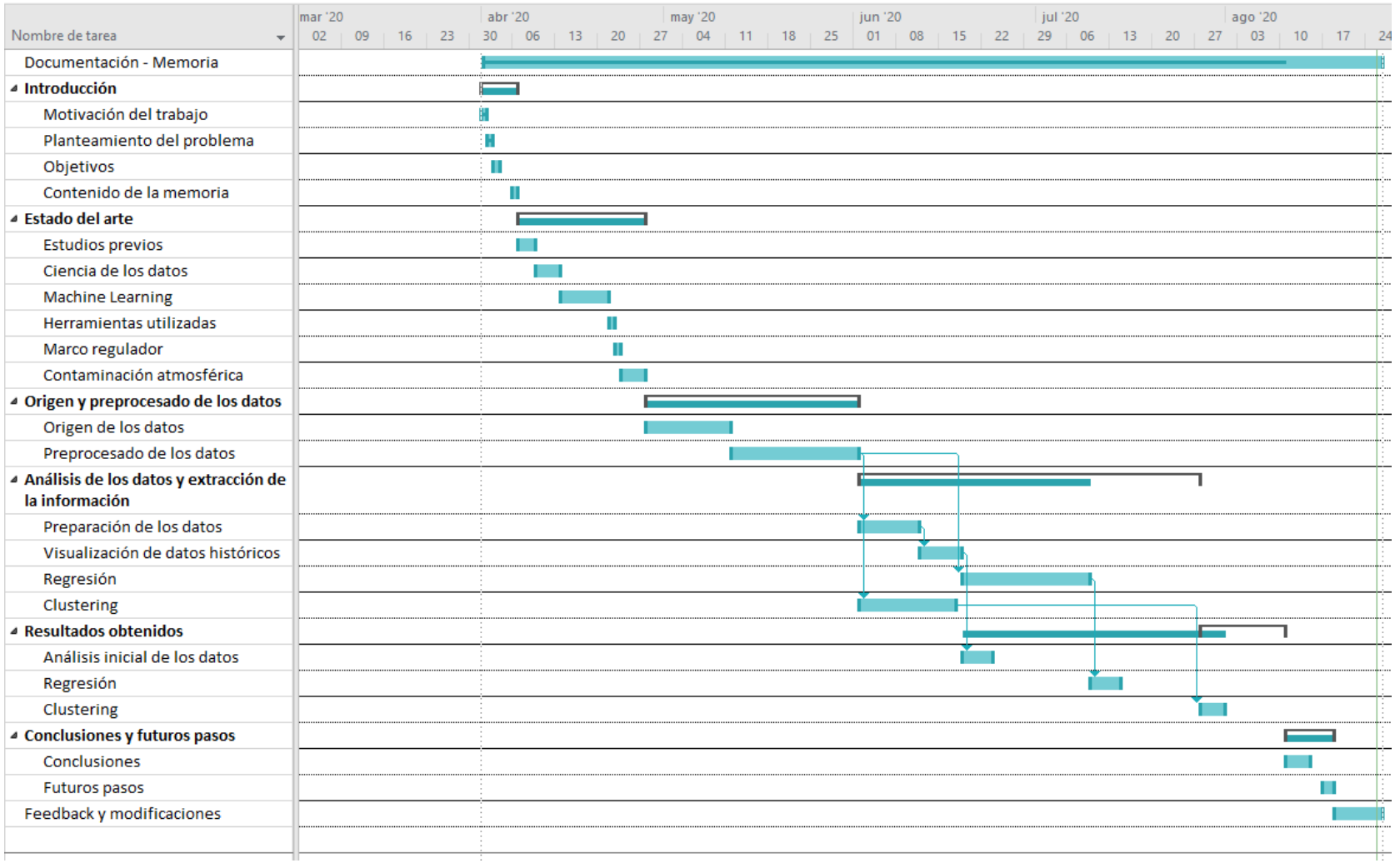


Figura 60. Diagrama de Gantt

En la Figura 60, además de la duración de cada una de las secciones, se ponen de manifiesto las dependencias entre secciones. Por ejemplo, el análisis de los datos no es posible sin haber obtenido previamente dichos datos, como es evidente.

Como se ha explicado anteriormente, el diagrama de Gantt se basa en el número de horas dedicadas al proyecto.

Documentación - Memoria	838 horas	106 días	mié 01/04/20	mié 26/08/20
Introducción	**32 horas**	**4 días**	**mié 01/04/20**	**lun 06/04/20**
Motivación del trabajo	8 horas	1 día	**mié 01/04/20**	mié 01/04/20
Planteamiento del problema	8 horas	1 día	**jue 02/04/20**	jue 02/04/20
Objetivos	8 horas	1 día	**vie 03/04/20**	**vie 03/04/20**
Contenido de la memoria	8 horas	1 día	lun 06/04/20	lun 06/04/20
Estado del arte	**120 horas**	**15 días**	mar 07/04/20	**lun 27/04/20**
Estudios previos	24 horas	3 días	mar 07/04/20	jue 09/04/20
Ciencia de los datos	16 horas	2 días	vie 10/04/20	lun 13/04/20
Machine Learning	48 horas	6 días	mar 14/04/20	mar 21/04/20
Herramientas utilizadas	8 horas	1 día	mié 22/04/20	mié 22/04/20
Marco regulador	8 horas	1 día	jue 23/04/20	jue 23/04/20
Contaminación atmosférica	16 horas	2 días	vie 24/04/20	lun 27/04/20
Origen y preprocesado de los datos	**200 horas**	**25 días**	mar 28/04/20	**lun 01/06/20**
Origen de los datos	80 horas	10 días	mar 28/04/20	lun 11/05/20
Preprocesado de los datos	120 horas	15 días	mar 12/05/20	lun 01/06/20
Análisis de los datos y extracción de la información	**320 horas**	**40 días**	**mar 02/06/20**	**lun 27/07/20**
Preparación de los datos	64 horas	8 días	mar 02/06/20	jue 11/06/20
Visualización de datos históricos	40 horas	5 días	jue 11/06/20	jue 18/06/20
Regresión	120 horas	15 días	vie 19/06/20	jue 09/07/20
Clustering	96 horas	12 días	lun 01/06/20	mié 17/06/20
Resultados obtenidos	**86 horas**	**10 días**	**mar 28/07/20**	**lun 10/08/20**
Análisis inicial de los datos	24 horas	3 días	jue 18/06/20	mar 23/06/20
Regresión	24 horas	3 días	jue 09/07/20	mar 14/07/20
Clustering	38 horas	4 días	lun 27/07/20	vie 31/07/20
Conclusiones y futuros pasos	**32 horas**	**6 días**	**mar 11/08/20**	**mar 18/08/20**
Conclusiones	16 horas	4 días	lun 10/08/20	vie 14/08/20
Futuros pasos	16 horas	2 días	lun 17/08/20	mar 18/08/20
Feedback y modificaciones	48 horas	6 días	mié 19/08/20	mié 26/08/20

Figura 61. Desglose de esfuerzo del proyecto

Como se puede comprobar en la Figura 61, los capítulos que han requerido un mayor esfuerzo son:

- **Origen y preprocesado de los datos**: gran parte del esfuerzo dedicado a esta sección se debe a la manipulación de los datos de contaminación, pues suponen una enorme cantidad de datos en un formato muy alejado del resultado deseado.

- Análisis de los datos y extracción de la información: la variedad de las técnicas y algoritmos aplicados ha supuesto un esfuerzo importante, ya no solo en la implementación, sino en la comprensión de estos. Son los resultados de este capítulo los utilizados en las secciones posteriores, por lo que su calidad y claridad son de gran importancia.

2. Presupuesto del proyecto

Los gatos asociados al proyecto se dividen en costes de hardware, costes de software y costes de personal o Recursos Humanos.

2.1. Hardware

Los costes de hardware incluyen la adquisición de los equipos necesarios para la realización de este proyecto.

Al utilizar un entorno web para el desarrollo que se encarga del coste computacional, la única funcionalidad necesaria que debe tener el equipo es acceso a Internet y cualquier tipo de ordenador actual tiene esta capacidad.

Se tiene en consideración que un equipo electrónico tiene un tiempo máximo de amortización de 120 meses (10 años) [41], y la duración del proyecto es de 106 dias, lo que resulta en aproximadamente 5,3 meses de duración teniendo en cuenta que un mes tiene 20 dias laborables.

Concepto	**Precio (€)**	**Periodo de amortización máximo (meses)**	**Uso (Meses)**	**Coste (€)**
Ordenador portátil	399	120	5,3	17,62

Tabla 22. Desglose de gastos asociados al hardware

2.2. Software

El único software utilizado, aparte de las librerías de Python, es el entorno de programación Google Colaboratory y el almacenamiento en la nube Google Drive, ambos gratuitos, por lo que el coste de software es 0€.

Concepto	**Precio (€)**
Google Drive	0
Google Colaboratory (Entorno de programación)	0

Tabla 23. Desglose de gastos asociados al software

2.3. Recursos humanos

Los recursos humanos necesarios para la realización de este proyecto es un Científico de Datos de categoría Junior. Se tiene en cuenta que el sueldo anual bruto de un cargo de este tipo es de 25000€ brutos al año y que el convenio laboral estipula que el máximo de horas trabajadas en un año es de aproximadamente 1800 horas, variando según el convenio [42].

Concepto	**Precio por hora (€)**	**Horas necesarias**	**Total (€)**
Científico de Datos Junior	13,89	838	11639,82

Tabla 24. Desglose de gastos asociados a los Recursos Humanos

Teniendo en cuenta el número de horas dedicadas a partir de la Figura 61 y el coste por hora de un trabajador con el perfil especificado, el coste total de Recursos Humanos es 11639,82 €.

2.4. Costes totales

Teniendo en cuenta los gastos previamente considerados, se calcula el total de costes asociado a la realización del presente proyecto.

Concepto	**Coste (€)**
Ordenador portátil	17,62
Google Drive	0
Google Colaboratory	0
Científico de Datos Junior	11639,82
Total	**11656,62**

Tabla 25. Desglose de gastos totales

8. REFERENCIAS

[1] WORLD HEALTH ORGANIZATION, W., 2020. *How Air Pollution is Destroying our Health.* 2020, Available from: https://www.who.int/news-room/spotlight/how-air-pollution-is-destroying-our-health.

[2] EUROPEAN ENVIRONMENT AGENCY, E., 2020. *Health Impacts of Air Pollution.* Jan 28, 2020, Available from: https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution.

[3] United Nations., 2015. *El Acuerdo De París.* París: Dec, 2015, Available from: https://unfccc.int/es/process-and-meetings/the-paris-agreement/el-acuerdo-de-paris.

[4] CUMMINGS, I., 2012. *Hean and Ozone can Produce Symptoms Mimicing Allergies.* August 10, 2012, Available from: https://weather.com/health/allergy/news/heat-ozone-mimic-allergies-20120810.

[5] FMI, F.M.I., s.f.. *How does Weather Affect Air Pollution?* sin fecha, Available from: https://en.ilmatieteenlaitos.fi/weather-and-air-quality.

[6] National Weather Service., 2020. *Clearing the Air on Weather and Air Quality.* 2020, Available from: https://www.weather.gov/wrn/summer-article-clearing-the-air.

[7] Waikato., 2020. *Direct Effects of Weather on Air Quality.* 2020, Available from: https://www.waikatoregion.govt.nz/environment/air/weather-and-air/#:~:text=For%20example%2C%20sunshine%2C%20rain%2C,chemical%20reactions%20in%20the%20air.

[8] M. De Sario, K.Katsouyanni, P. Michelozzi., 2013. *Climate Change, Extreme Weather Events, Air Pollution and Respiratory Health in Europe.* Available from: https://erj.ersjournals.com/content/42/3/826

[9] NIÑO, M., 2016. *Repaso De Diferentes Perspectivas Para Entender La "Ciencia De Datos" (Data Science).* Nov 11, 2016, Available from: http://www.mikelnino.com/2016/11/repaso-perspectivas-ciencia-datos-data-science.html.

[10] GRANVILLE, V., 2017. *Types of Machine Learning Algorithms in One Picture.* July 27, 2017, Available from: https://www.datasciencecentral.com/profiles/blogs/types-of-machine-learning-algorithms-in-one-picture.

[11] HOLLANDER, G., 2019. *What is Structured Data Vs. Unstructured Data?* September 26, 2019, Available from: https://www.m-files.com/blog/what-is-structured-data-vs-unstructured-data/.

[12] PROGRAMADOR, I., 2019. *Clusterin Jerárquico Con Python y Scikit-Learn.* July 9, 2019, Available from: https://www.instintoprogramador.com.mx/2019/07/clustering-jerarquico-con-python-y.html.

[13] Wikipedia., 2020. *Regresión Lineal.* July 27, 2020, Available from: https://es.wikipedia.org/wiki/Regresión_lineal.

[14] JavaTPoint., s.f.. *Classification Algorithm in Machine Learning.* sin fecha, Available from: https://www.javatpoint.com/classification-algorithm-in-machine-learning.

[15] lucidar.me., 2020. *Simplest Perceptron.* May 3, 2020, Available from: https://lucidar.me/en/neural-networks/simplest-perceptron/.

[16] Wikipedia., 2020. *Perceptrón Multicapa.* June 12, 2020, Available from: https://es.wikipedia.org/wiki/Perceptrón_multicapa.

[17] alexis96., unknown. *Red Neuronal Recurrente.* unknown, Available from: https://alexis96.github.io/proyecto-RNN/.

[18] OLAH, C., 2015. *Understanding LSTM Networks.* August 27, 2015, Available from: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[19] INDIA, S., 2020. *Top 15 Open-Source Data Science Tools to Learn in 2020.* January 30, 2020, Available from: https://in.springboard.com/blog/top-open-source-data-science-tools/.

[20] HAYES, B., 2019. *Programming Languages most used and Recommended by Data Scientists.* January 13, 2019, Available from:

https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/.
[21] Google Colab., 2020. *Colaboratory.* 2020, Available from: https://colab.research.google.com/notebooks/intro.ipynb.
[22] Ayuntamiento de Madrid., 2018. *Protocolo De Actuación Para Episodios De Contaminación Por Dióxido De Nitrógeno De La Ciudad De Madrid.* Protocolo ed. Medio ambiente y movilidad, October 9, 2018,.
[23] Ayuntamiento de Madrid., 2020. *Conjunto De Datos Destacados (Madrid).* 2020, Available from: https://datos.madrid.es/portal/site/egob/.
[24] Agencia Estatal de Meteorología., 2020. *Aemet Opendata.* August 15, 2020, Available from: https://opendata.aemet.es/centrodedescargas/inicio.
[25] Ayuntamiento de Madrid., s.f.. *Intérprete De Ficheros De Datos Diarios y Tiempo Real.* sin fecha, Available from: https://datos.madrid.es/FWProjects/egob/Catalogo/MedioAmbiente/Aire/Ficheros/Interprete_ficheros_%20calidad_%20del_%20aire_global.pdf.
[26] European Commission., 2019. *Air Quality Standards.* December 31, 2019, Available from: https://ec.europa.eu/environment/air/quality/standards.htm.
[27] ANALYTICA, V.m., unknown. *Normal Distribution.* unknown, Available from: https://wiki.analytica.com/index.php?title=Normal_distribution.
[28] BOWNLEE, J., 2019. *Loss and Loss Functions for Training Deep Learning Neural Networks.* October 23, 2019, Available from: https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/.
[29] BONAROS, B., 2019. *K-Means Elbow Method Code for Python.* August 12, 2019, Available from: https://predictivehacks.com/k-means-elbow-method-code-for-python/.
[30] Wikipedia., 2020. *Silhouette (Clustering).* April 10, 2020, Available from: https://es.wikipedia.org/wiki/Silhouette_(clustering).
[31] Junta de Andalucía., s.f.. *Emisiones De Gases Precursores Del Ozono Troposférico.* sin fecha, Available from: http://www.juntadeandalucia.es/medioambiente/site/portalweb/menuitem.7e1cf46ddf59bb227a9ebe205510e1ca/?vgnextoid=bd594e4b4b836110VgnVCM1000000624e50aRCRD&vgnextchannel=66c8445a0b5f4310VgnVCM2000000624e50aRCRD&lr=lang_es#:~:text=Se%20forma%20a%20través%20de,monóxido%20de%20carbono%20(CO).
[32] VAN DER MAATEN, L. and HINTON, G., 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, November, 2008, no. 9, pp. 2579-2605.
[33] PIATETSKY, G., 2020. *Python Leads the 11 Top Data Science, Machine Learning Platforms: Trends and Analysis.* Available from: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html.
[34] Python TM., 2013. *Guía De Buenas Prácticas De Python.* August 1, 2013, Available from: https://www.python.org/dev/peps/pep-0008/.
[35] IQ Air Group., 2020. *2019 World Air Quality Report.* February 17, 2020, Available from: https://www.iqair.com/world-most-polluted-cities.
[36] JOLLIFFE, I.T. and CADIMA, J., 2016. **Principal Component Analysis: A Review and Recent Developments**. 2065th ed., 13/04/2016, April 13, 2016, vol. 374 ISSN 1471-2962. DOI https://doi.org/10.1098/rsta.2015.0202.
[37] SUSSILLO, D. and ABBOTT, L.F., 2015. ***Random Walk Initialization for Training very Deep Feedforward Networks***. February 27, 2015, Available from: https://arxiv.org/abs/1412.6558.
[38] Agrawal, R.; Srikant, R. *Mining sequential patterns. In Proceedings of the Eleventh International Conference on Data Engineering*, Taipei, Taiwan, 6–10 March 1996.
[39] Zaki, M.J. SPADE: An efficient algorithm for mining frequent sequences. *Mach. Learn.* **2001**, *42*, 31–60. [**Google Scholar**] [**CrossRef**]
[40] Ayres, J.; Flannick, J.; Gehrke, J.; Yiu, T. Sequential pattern mining using a bitmap representation. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002.
[41] Agencia Tributaria., 2020. Tabla De Amortización Simplificada. April 6, 2020, Available from: https://www.agenciatributaria.es/AEAT.internet/Inicio/Ayuda/Manuales__Folleto

s_y_Videos/Manuales_practicos/_Ayuda_Folleto_Actividades_economicas/3__Impuesto_sobre_la_Renta_de_las_Personas_Fisicas/3_5_Estimacion_directa_simplificada/3_5_4__Tabla_de_amortizacion_simplificada/3_5_4__Tabla_de_amortizacion_simplificada.html.

[42] Fernando., 2020. Horas de trabajo anuales. ¿Cómo calcularlas?. May 15, 2020, Available from: https://asesorias.com/empresas/normativas/laboral/jornada/horas-trabajo-anuales

ANEXO I: SUMMARY

1. INTRODUCTION

1.1. Motivation

Nowadays, the impact of human beings on nature is an increasingly serious reality that can occur in many different ways. Air pollution in particular is one of the biggest concerns because of the way it affects people's health, causing diseases, from asthma to cancer in the most extreme cases. The World Health Organization estimates that 9 out of 10 people in the world breathe polluted air, resulting in a loss of 7 million lives per year [1].

For this reason, a large number of government agencies, and universities have published studies on the impact of these pollutants on health and their evolution over time, such as this study by the European Union that analyzes the levels of contamination in various countries and the number related deaths [2].

This is not the only problem associated with air pollution, since carbon dioxide (CO2) is the main responsible for climate change along with methane (CH4) and nitrous oxide (N2O), whose levels have increased worryingly in the past few years. In the case of CO2, the most abundant one, at the time of the Industrial Revolution had levels of 280 parts per million, currently reaching 415 ppm, a figure that can be verified in real time through the official website of NASA https://climate.nasa.gov.

To control this situation and reduce the emissions of these pollutants, in December 2015 the Paris Agreement [3] was signed, a global treaty to limit the effects of climate change and to try to avoid the global temperature rise by 1.5 ºC.

The aim of this document is to study how the emission of different pollutants in the city of Madrid has evolved and to check if they bear any relationship between them or with meteorological factors, applying the most popular techniques and algorithms for data analysis and artificial intelligence.

1.2. Problem statement

The problem statement is explained below, at what point it starts and what are the techniques that are applied.

Starting from the problems associated with air pollution presented in the previous section, it is possible to use Data Science techniques, both to expand the expand the knowledge

that I currently have by analyzing historical data of different magnitudes, and to be able to predict future contamination scenarios and to act accordingly before they happen.

The application of these Data Science techniques is only possible starting from the available data, and in this work real data of the city of Madrid data have been obtained from various official sources, so it is necessary to unify them in a suitable format for their analysis and then, the study is carried out and the necessary techniques can be applied.

This study is based on the use of these techniques for the analysis of historical data of meteorological data and air pollution for the extraction of information and behaviors followed by the use of Machine Learning algorithms to verify their scope and usefulness in this kind of problem.

1.3. Goals

The main goal of this paper is to illustrate how data analysis and artificial intelligence (and, ultimately, technology) can be used for a social purpose, such as a study of the quality of the air, in the same way which it is being applied increasingly in other areas such as medicine, when processing radiographs and diagnose diseases, or in the diagnosis of outbreaks of fires in satellite imagery.

To sum up, the main goals are:

I. Show what data is available to the public, which institutions are offering them, and what useful information can be obtained from it.

II. Use the most popular tools today for data analysis, exploration, and representation.

III. Illustrate, through visualization tools, a history of pollution and meteorological levels and what the current trend is.

IV. Gain knowledge and fluency in the Python programming language, the most widely used (along with R) for data analysis and artificial intelligence.

V. Try to expand the pollution criteria of the city of Madrid by discovering patterns in the data obtained.

VI. Explore the precision, scope, and usefulness of the most widely used Machine Learning algorithms, both unsupervised and supervised.

VII. Use techniques for predicting future values to anticipate the levels of contaminants that may occur and thus apply contamination protocols before they occur.

2. OBTAINED RESULTS

The results obtained after analyzing the data and the application of the techniques explained during the project will be discussed below.

2.1. Initial data analysis

On one hand, analysis of historical data is interpretable, rather than a technological way, from a point of view focused on the problems previously presented in section 7.1, as it is where is put into context the role of air pollutants.

Weather conditions remain stable throughout the time frame referred to in this study, although it would be necessary to carry out a much larger study in order to draw reliable conclusions.

On the other hand, a large majority of pollutants have seen their levels reduced in recent years, especially SO2 and NO2 have suffered a very pronounced reduction since 2010. However, O3 levels have increased dramatically, which it can seriously affect people's health [31].

2.2. Regression

The main approach to predict future values has been to apply Deep Learning, varying the design of the neural network, while maintaining the hyperparameters to compare the results with higher reliability. The ARIMA model has been used to compare the results between the classical prediction methods and the current machine learning algorithms, although the data on which the error has been calculated is a subset of the data used with neural networks.

Model	**Error (MSE)**
ARIMA	6.10229770762e-07
simple univariate LSTM	2.27970209444e-06
Simple multivariate LSTM	5.78301387732e-07
Multivariate multilayer LSTM	1.19993757717e-06

Table 26. Error of the regression models

It is interesting to see how the error of the most complex model of all (multivariate multilayer LSTM) is not the lowest of all. This is because a larger neural network does not necessarily imply that its operation will be better, since the design of the model must adjust to the characteristics of the data involved and it is necessary to perform multiple experiments with different combinations of neural networks and hyperparameters.

Although the ARIMA model error is less than some Deep Learning models error, remember that the data on which the model has been adjusted are those corresponding to a single year. Although, this proves that the classic regression techniques can compete with much more modern and complex algorithms, and it can be considered in the resolution of problems.

2.3. Clustering

Although all observations have been grouped into their corresponding clusters by several algorithms, it is advisable to seek a way to graphically represent these groups to have an overview.

Fecha	T-Media	Precip	Vel-Viento	Pres-Media	Escenario	SO2	NO2	PM10	O3	PM2.5	Escenario	K-Means	Hierach_3	Hirach_4
2003-01-01 01:00:00	-0.964198	1.116283	0.262839	-1.097426	0	0.295134	0.254113	0.113652	-0.551432	-0.296567	0	1	0	0

Table 27. Data sample after applying clustering algorithms

However, this grouping has been done by the representation of the data obtained from temperature, rainfall, wind speed, atmospheric pressure, SO2, NO2, PM10, O3 and PM2.5. That is, they consist in 9 dimensions and it is impossible to graphically plot them.

At this point, there are two possible options to study the results of clustering: study the clusters in each of the variables or apply dimensionality reduction techniques to represent all the possible information in a two-dimensional space.

2.3.1. Study of each of the magnitudes

It is intended, by carrying out the study of each magnitude separately, to identify patterns in the clusters associated to the data through each of the variables.

The first step is to check the historical data, distinguishing the different clusters of each magnitude. Figure 62 shows the data for each of the variables over time, differentiating the cluster associated with each point with a different color.

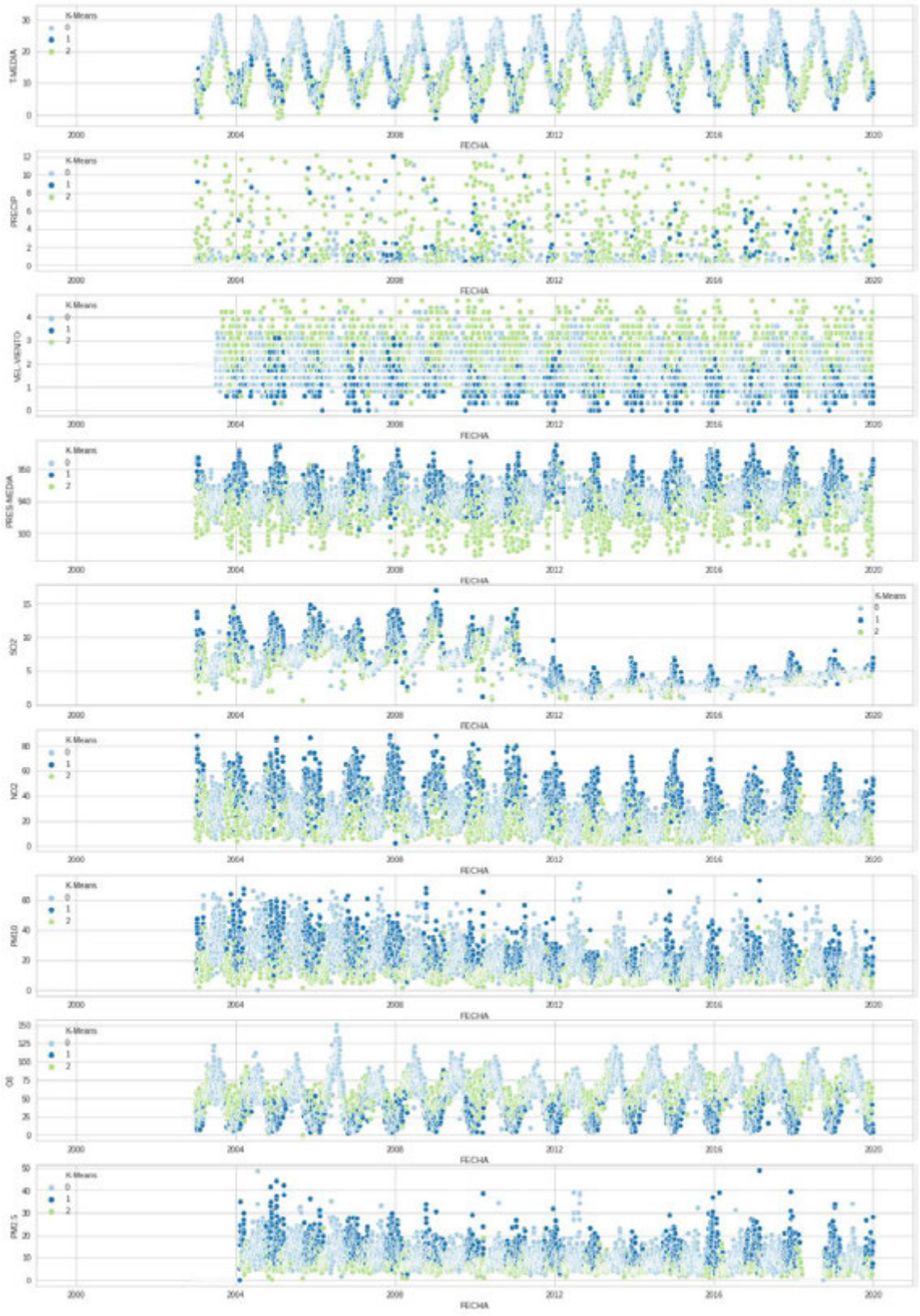


Figure 62. Particular view of the K-Means results

The results of the 3 clustering algorithms follow a very similar behavior when classifying the data, so the K-Means results will be used to illustrate the criteria since they seem quite generalized.

The first conclusion reached by observing Figure 62 is that the criteria are repeated every year, so it is redundant to observe the whole set. One approach may be to display data in a particular year (2019, for example) and marking the end of each season (in order are the end of winter, late spring, late fall, and winter again).

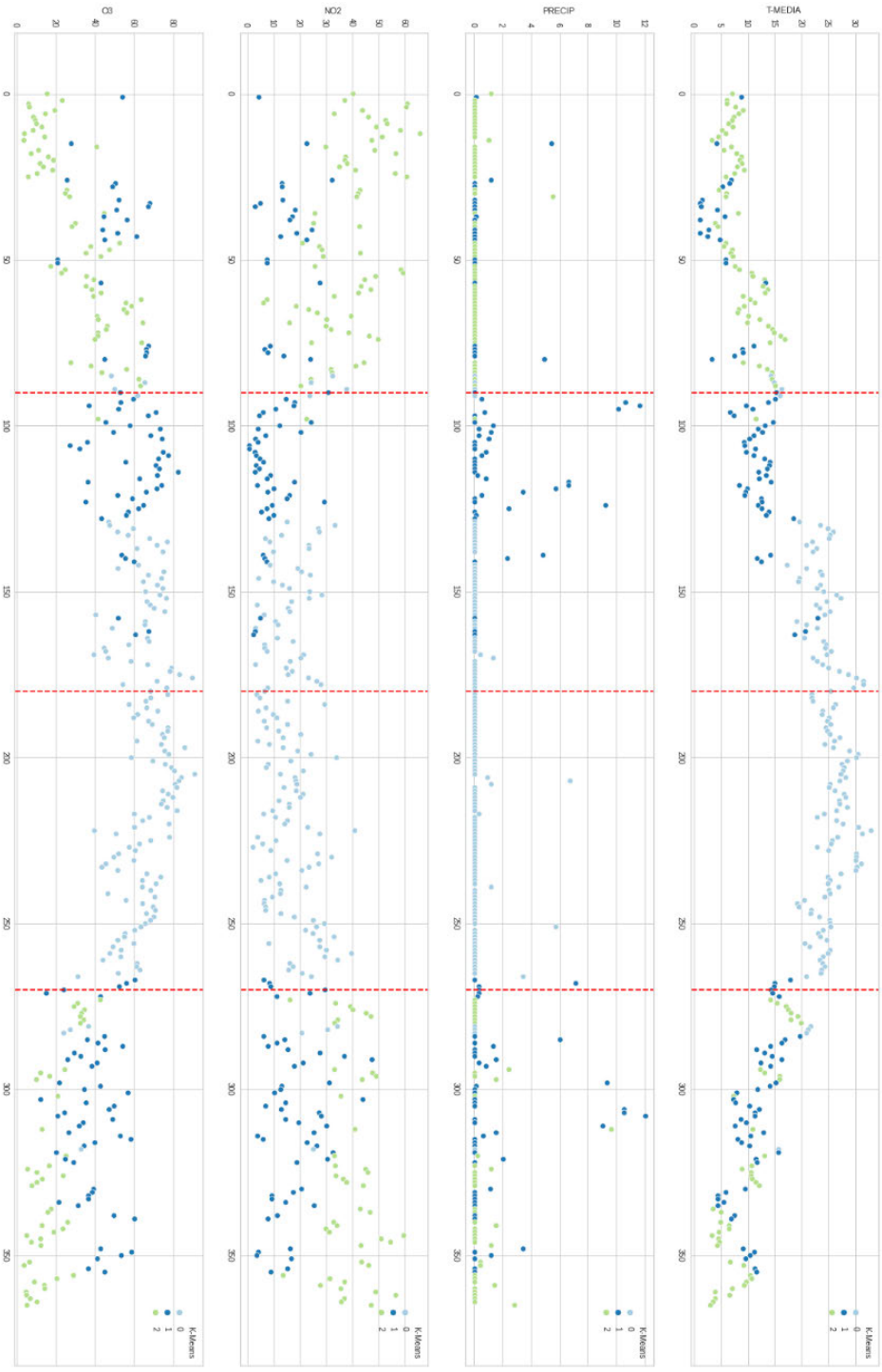


Figure 63. K-Mean results differentiating seasons of the year

It is obvious in Figure 63 that there is a very marked time component when classifying the data in the different clusters. For example, summer season data belonging (almost entirely) to the cluster 0 and winter to cluster 2, dividing the fall and spring in cluster 1.

This seasonal approach seems to have been the main criterion when grouping the data, which makes sense considering that the weather conditions are very marked by the seasons and directly influence on pollutants.

2.3.2. Dimensionality reduction

There are techniques for reducing the number of dimensions in the data while maintaining the information, such as Principal Component Analysis (or PCA) and t-SNE (*t-distributed Stochastic Neighbor Embedding*).

The aim of representing data using this type of algorithms is to evaluate how data is grouped by similarity and, on this basis, to evaluate how the clustering algorithms have acted and whether the groups are clearly differentiated.

After applying each of the algorithms, we have two dimensions for each one that we could call the X coordinate axis and the Y coordinate axis. Using these coordinates axis, the data is shown in four graphs:

1. Membership of the data divided into groups by the K-Means algorithm.

2. Identification of anti-pollution scenarios. It has been decided to represent this category in case some clustering algorithm would have separated the most serious scenarios in the same group or cluster.

3. Membership of the data divided into three groups by hierarchical clustering.

4. Membership of the data divided into four groups by hierarchical clustering.

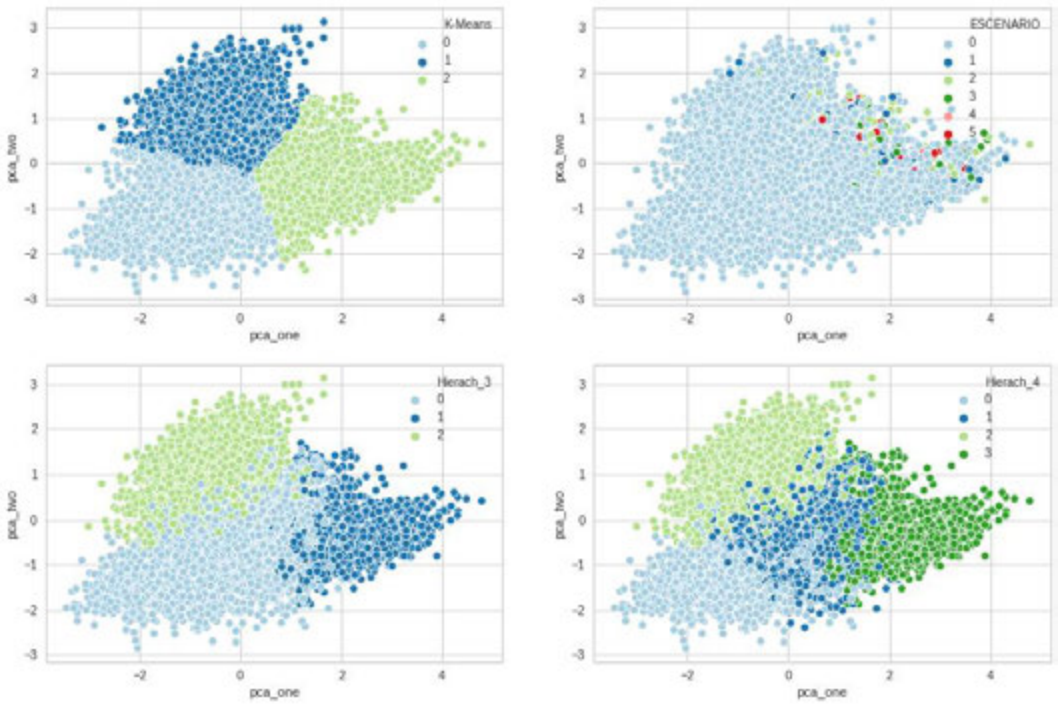


Figure 64. 2D representation of the clustering results using PCA

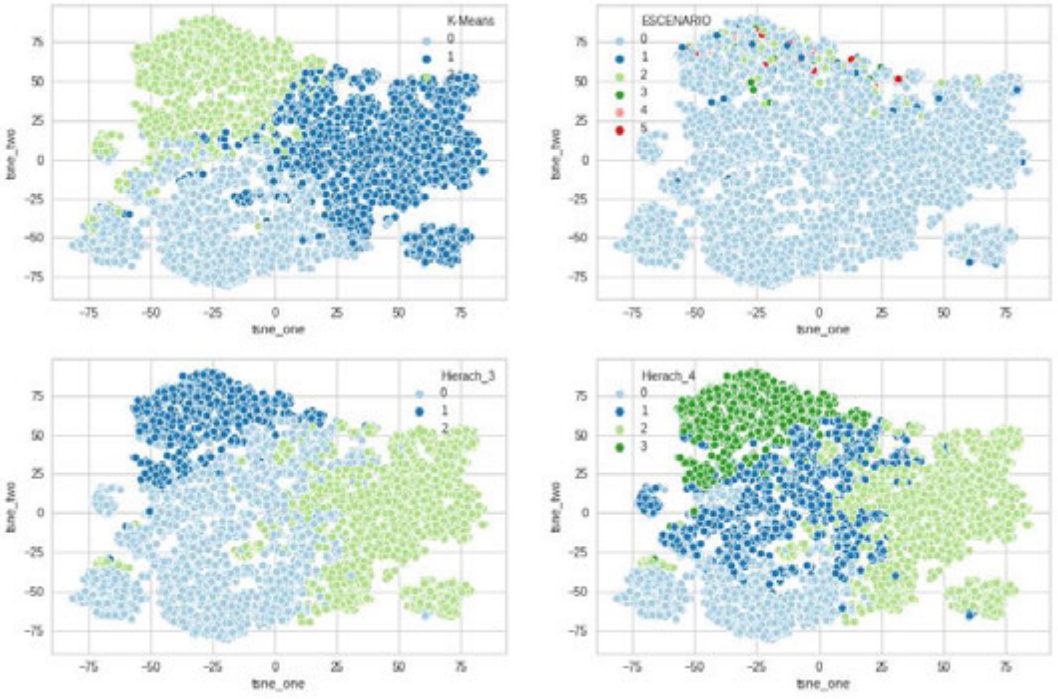


Figure 65. 2D representation of the clustering results using t-SNE

Performing a visual comparison between the results of the PCA and t-SNE (Figure 64 and Figura 65, respectively) it can be seen that with t-SNE, as a general rule, most similar data are grouped together as they are separated from the rest, and in the end, there are more differentiated groups.

Although both algorithms fulfill the objective of representing data in a reduced dimensional space, it is clear that t-SNE is widely used in data analysis and science for data visualization in a 2 or 3 dimensional space [32].

Another argument against PCA is that the amount of information that it can represent varies depending on the number of the chosen components.

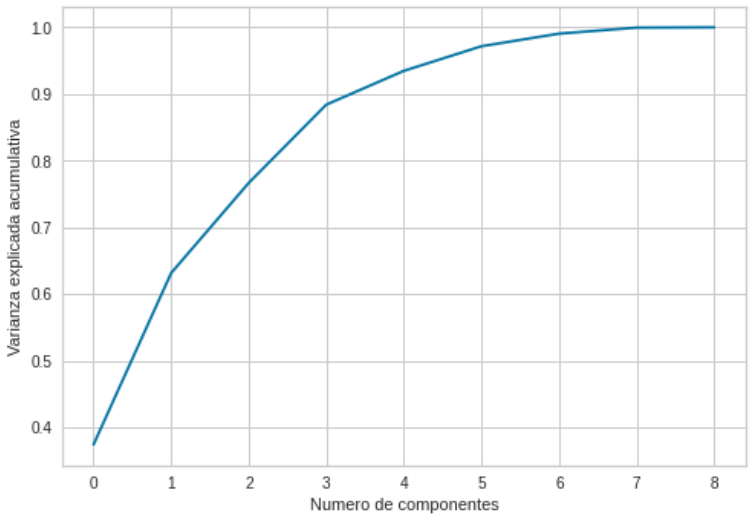


Figure 66. Cumulative explained variance

Figure 66 shows that, with the data we have available, with two components approximately 77% of the data is represented, which is not a good quantity since 90% or more is usually acceptable, so the information represented with PCA results should not be considered representative, but merely illustrative.

3. CONCLUSIONS AND FUTURE WORK

3.1. Conclusions

In view of the results obtained and based on the goals defined at the beginning of this document, the following conclusions have been found:

I. The available data to perform a study of weather conditions and air pollution are accessible to anyone since the prerequisites are minimal:

Requirement	**Rate**
Knowledge on HTTP requests	Weather data is accessible via HTTP requests to the AEMET' API REST.
Email	Required to receive the AEMET API key to request atmospheric data by HTTP requests.
Internet access	Internet access is essential to download air pollution data from the Madrid City Council website and to obtain meteorological data.

Table 28. Requirements to obtain weather data

The data obtained is sufficient to meet the objective of this project, and the necessary steps have been carried out to unify the original data in a more suitable format for study.

II. During the course of this project, practically all the modifications that have been made to the data and its visualization have been carried out using the *pandas* library, one of the most used free data analysis tools in the present.

Software	2019 % share	2018 % share	2017 % share
Python	65.8%	65.6%	59.0%
RapidMiner	51.2%	52.7%	31.9%
R Language	46.6%	48.5%	56.6%
Excel	34.8%	39.1%	31.5%
Anaconda	33.9%	33.4%	24.3%
SQL Language	32.8%	39.6%	39.2%
Tensorflow	31.7%	29.9%	22.7%
Keras	26.6%	22.2%	10.7%
scikit-learn	25.5%	24.4%	21.9%
Tableau	22.1%	26.4%	21.8%
Apache Spark	21.0%	21.5%	25.5%

Table 29. Most used data Science / Analysis / Machine Learning software [33]

In addition to *pandas*, libraries that are among the most used nowadays have been used throughout the project, such as *Keras* for neural network design and *scikit-learn* for unsupervised learning algorithms implementation.

III. The historical data of each of the magnitudes, both meteorological and atmospheric pollution, have been analyzed and visualized through graphs with the entire timeline. And, in cases where it has been necessary, a specific time range has been specified.

For example, the magnitude of PM2.5 only began to be registered since the year 2005, so it required a specific treatment over others.

In case that the data show a seasonal behavior, the necessary tools have been used to decompose them into seasonal components (observed data, trend, seasonality, residual) in order to study the trend over a defined period of time.

Finally, the usefulness and implications of the linear correlation coefficient when analyzing the linear relationships between the different magnitudes have been demonstrated.

IV. Thanks to this project, it has been possible to achieve considerable fluency in handling the tools used, and in the Python programming language.

It is remarkable the interaction that we discovered that exists between the different libraries that we used for this project. For example, *pandas library* functions to visualize data from their own data frames use their own visualization libraries as backend to achieve a more descriptive style; or *numpy*, a very extensive library of mathematical functions can be used to prepare the input data to a network neural, or manipulate data used by *pandas library*.

Python mechanisms have also been understood for code structure, to make it illustrative, using descriptive function names and comments following best practices [34].

V. Clustering algorithms have been used in order to identify patterns in the available data, which allow us to define pollution scenarios, and thus expand the existing ones in the city of Madrid.

This objective has not been achieved since the result of these algorithms has turned out to be the grouping of the data based on the season of the year to which they belong, to a greater or lesser extent. This shows that, although this is not the desired result, the algorithms have successfully identified the similarities between the data, since it has been possible to verify that a large part of the magnitudes have a common behavior throughout each year.

VI. Various Machine Learning algorithms and techniques with different applications and objectives have been used. For each of them, the objective of their implementation and the results obtained have been explained, in words and graphically.

The results of the unsupervised learning algorithms, being unable to measure quality in an objective way, have been analyzed in different ways and evaluated subjectively and in accordance with the goals set before carrying out this project. The supervised learning algorithms have been compared between them using the error of their predictions with the real data and their learning process has been explained to understand how we have reached these conclusions.

VII. Regression algorithms have proven to be very accurate in generating future data values. Being viable data and on the pollutant that determines the pollution scenarios, it is possible to predict the data from each of the city's stations (since it is necessary to take into account the different stations in the different urban areas) to calculate the criteria that they would be applied in these supposed future data and, in this way, act proactively before reaching limit levels of contamination if necessary.

3.2. Future work

The future work to expand the information in this study can be separated into many sections:

- Expansion of the data used with population data in the city of Madrid, traffic data in the hours at which the pollutant measurements were made and energy expenditure associated with different fuels.

- Apply more advanced clustering techniques that allow data mining by considering the sequential component of the data and not as separate data samples. Some of these algorithms are Generalized Sequential Pattern (GSP) [38], Sequential Pattern Discovery using Equivalence classes (SPADE) [39] and Sequential Pattern Mining (SPAM) [40].

- Define an architecture and parameterization criteria of the neural networks for maximizing de quality of predictions

- Implement more complex and recent neural networks than LSTMs such as GRU or Transformer, also oriented to time series regression.

- With the aim of expanding the knowledge of Python, create a dashboard hosted on the web where the findings of this study can be consulted using the libraries Dash and Plotly for the creation of interactive graphics.

- Expand this study with the relationship between the different measurement stations, possibly based on the distance and location of each one.